UNIVERSITY OF CAPE TOWN

DEPARTMENT OF COMPUTER SCIENCE

# CS/IT Honours
# Final Paper 2021

Title: Propositional Defeasible Explanation

Author: Lloyd Everett

Project Abbreviation: PDE

Supervisor(s): Tommie Meyer

| **Category** | *Min* | *Max* | Chosen |
|---|---|---|---|
| Requirement Analysis and Design | *0* | *20* | 0 |
| Theoretical Analysis | *0* | *25* | 25 |
| Experiment Design and Execution | *0* | *20* | 0 |
| System Development and Implementation | *0* | *20* | 0 |
| Results, Findings and Conclusions | *10* | *20* | 20 |
| Aim Formulation and Background Work | *10* | *15* | 15 |
| Quality of Paper Writing and Presentation | *10* | | 10 |
| Quality of Deliverables | *10* | | 10 |
| Overall General Project Evaluation (*this section allowed only with motivation letter from supervisor*) | *0* | *10* | 0 |
| **Total marks** | | **80** | |

# Explanation for KLM-Style Defeasible Reasoning

Lloyd Everett
University of Cape Town
Cape Town, South Africa
lloydeverett@gmail.com

## ABSTRACT

Explanation services are a crucial aspect of symbolic reasoning systems but they have not been explored in detail for defeasible formalisms such as KLM. We evaluate prior work on the topic with a focus on KLM propositional logic and find that a form of defeasible explanation initially described for Rational Closure which we term *weak justification* represents a convincing defeasible analogue to classical justification. We characterise this notion of explanation declaratively by adapting the postulates for LM-rationality and extend its definition and algorithm to the case of Relevant Closure. These results illustrate that weak justification obeys several intuitive properties, that it extends classical justification in much the same way that defeasible entailment extends classical entailment, and that weak justification can successfully be adapted to defeasible entailment formalisms inferentially stronger than Rational Closure. The properties we give for weak justification may also offer a perspective from which to evaluate other notions of defeasible explanation.

## CCS CONCEPTS

• **Computing methodologies → Nonmonotonic, default reasoning and belief revision**; *Causal reasoning and diagnostics*; • **Theory of computation** → *Automated reasoning*.

## KEYWORDS

knowledge representation and reasoning, explainable artificial intelligence, defeasible reasoning, Rational Closure, Relevant Closure

## 1 INTRODUCTION

Explanation services indicate to users of symbolic reasoning systems which parts of their knowledge base lead to particular conclusions. This is helpful particularly when the reasoner is giving unexpected results since it allows the user to identify the culprit knowledge base statements and thus debug their knowledge base [12]. Explanation services have also been found to improve knowledge base comprehension, particularly if the user is not familiar with the knowledge base [1], and to improve users' confidence in the reasoning system [2]. There is also some evidence that formalisms of explanation can be theoretical tools in their own right; for example, Casini et al. [4] base their work on Relevant Closure fundamentally on classical justification, a form of classical explanation.

Explanation has been explored in detail for classical logics and explanation services are offered in many reasoning systems based on classical logics [12, 15]. However, classical logics are known to be ineffective at modelling certain kinds of information. In particular, they are poor at modelling information that *typically* holds, but for which there might be exceptions. A quintessential example here attempts to model penguins and birds simultaneously in a classical logic: suppose we express the statements "birds fly", "penguins are birds" and "penguins do not fly" in a classical logic knowledge base. Since there is both a way to conclude that penguins fly and that penguins do not fly, classical logic entailment will conclude that penguins do not exist—a result that is clearly undesirable. The only way to correctly handle this situation in classical logic is to state every possible exception upfront, i.e. "birds fly, except for penguins", but this is unwieldy and impractical for larger knowledge bases.

A *defeasible* logic on the other hand will enable us to directly express that "birds *typically* fly" and when reasoning we will correctly identify that penguins, although they are birds, may have some exceptional properties not typically characteristic of birds. This brings us to the focus of our research. Although well-understood in the classical case, explanation has not yet been explored in detail for defeasible reasoning apart from some foundational work [3, 9]. Our work aims to improve our understanding of explanation for defeasible formalisms and where relevant to provide algorithms for the practical implementation of explanation services.

There are many approaches to defeasible reasoning, but a particularly compelling approach that has been studied at length in the literature [5, 6, 8, 17, 18] is the KLM approach suggested by Kraus, Lehmann and Magidor [14]. One of the major appeals of KLM is that it can be viewed from two different angles, each with its own advantages: either using a series of postulates asserting behaviours we intuitively expect of the defeasible reasoning formalism, or using a model-theoretic semantics perhaps not as obviously intuitive but more amenable to computation by means of reasoning algorithms. These two perspectives are linked by results in the literature [7, 11, 17]. The duality here is an important feature of our analysis and in fact one of our main results suggests that defeasible explanation can likewise be characterised using declarative properties.

Because it has a simpler semantics, and because it can be seen as a foundation for more complex logics such as the popular description logics, we will focus exclusively on propositional logic. However, we expect that many of the ideas expressed in this paper will have analogues in the case of description logic given the general successes in the literature of translating principles for KLM propositional logic to description logic [4, 5, 8, 18].

## 2 BACKGROUND

### 2.1 Classical Propositional Logic

We begin with a finite set $\mathcal{P} = \{p, q, \cdots\}$ of *propositional atoms*. The binary connectives $\wedge, \vee, \rightarrow, \leftrightarrow$ and the unary negation operator $\neg$ are used recursively to form propositional formulas such as $\neg(p \vee q) \rightarrow r$. The set of all such formulas over $\mathcal{P}$ is the *propositional language* $\mathcal{L}$.

An *valuation* is a function $\mathcal{P} \rightarrow \{\mathrm{T}, \mathrm{F}\}$ that assigns a truth value to each atom in $\mathcal{P}$. We say that a formula $A \in \mathcal{L}$ is *satisfied* by a valuation $\mathcal{I}$, written $\mathcal{I} \models A$, if $A$ evaluates to true according to the truth value assignments in $\mathcal{I}$ for atoms in $A$ and the semantics of the operators in $A$ which should be familiar from Boolean algebra. For example, if $\mathcal{I}(p) = \mathrm{T}$ and $\mathcal{I}(q) = \mathrm{F}$, then $\mathcal{I} \models p \vee q$ but $\mathcal{I} \not\models p \wedge q$. The valuations that satisfy a formula $A$ are referred to as *models* of $A$, and the set of models of $A$ is denoted $\mathrm{Mod}(A)$. Finally, we assert that $\top$ is a propositional formula satisfied by every interpretation and $\bot$ is a formula not satisfied by any interpretation.

A finite set of propositional formulas is a *knowledge base* $\mathcal{K}$. A valuation is a model of a knowledge base $\mathcal{K}$ if it is a model of every formula in $\mathcal{K}$, i.e. $\mathrm{Mod}(\mathcal{K}) = \bigcap \{\mathrm{Mod}(A) \mid A \in \mathcal{K}\}$. A knowledge base $\mathcal{K}$ *entails* a formula $A$, denoted $\mathcal{K} \models A$, if $\mathrm{Mod}(\mathcal{K}) \subseteq \mathrm{Mod}(A)$. A formula $A$ can also entail a formula $B$, denoted $A \models B$, if $\mathrm{Mod}(A) \subseteq \mathrm{Mod}(B)$.

*Example 2.1.* As a simple example of reasoning using classical propositional logic, suppose we have a knowledge base $\mathcal{K}$ of the following statements:

(1) Tweety is a bird ($t \rightarrow b$).
(2) Birds fly ($b \rightarrow f$).

Since $\mathrm{Mod}(\mathcal{K}) \subseteq \mathrm{Mod}(t \rightarrow f)$, we have $\mathcal{K} \models t \rightarrow f$. In other words, $\mathcal{K}$ entails that Tweety flies.

Classical logics have the property of *monotonicity*, which intuitively means that adding statements to a knowledge base cannot render entailments that were previously true false:

Proposition 2.2 (Classical Monotonicity). *For a knowledge base $\mathcal{K}$ and a propositional formula $A$, if $\mathcal{K} \models A$ then $\mathcal{K}' \models A$ for any superset knowledge base $\mathcal{K}' \supseteq \mathcal{K}$.*

## 2.2 Classical Justification

Justifications are perhaps the most common and basic form of explanation and more sophisticated explanation services usually rely on justifications as an elementary tool [9, 12]. We will find that classical justifications are central to our study of defeasible justification and Relevant Closure.

*Definition 2.3.* A knowledge base $\mathcal{J}$ is a *justification* for an entailment[1] $\mathcal{K} \models A$ if $\mathcal{J}$ is a subset $\mathcal{J} \subseteq \mathcal{K}$ such that $\mathcal{J} \models A$ and there is no proper subset $\mathcal{J}' \subset \mathcal{J}$ such that $\mathcal{J}' \models A$.

The latter condition—the requirement that there is no proper subset for which the entailment holds—is often referred to as the condition of being *minimal*, so that the justifications of $\mathcal{K} \models A$ are the *minimal subsets* of $\mathcal{K}$ that entail $A$. Note also that one entailment may have many justifications and therefore we introduce the following notation:

*Definition 2.4.* The set of all justifications for $\mathcal{K} \models A$ is denoted $\mathcal{J}(\mathcal{K}, A)$.

The utility of justifications is that they identify the statements in $\mathcal{K}$ that give rise to the entailemt $\mathcal{K} \models A$. The following example illustrates this idea.

[1]Strictly, it is incorrect to refer to $\mathcal{K} \models A$ as an "entailment": the statement $\mathcal{K} \models A$ is the assertion that the pair $(\mathcal{K}, A)$ belongs to the entailment relation $\models$ and it would be more correct to refer to the pair. However, we believe this usage is more intuitive and that the rigorous meaning should be clear from context.

**Table 1: Examples of classical justifications for Example 2.5**

| $A$ | $\mathcal{J}(\mathcal{K}, A)$ |
|---|---|
| $t \rightarrow b$ | $\{\{t \rightarrow b\}\}$ |
| $t \rightarrow w$ | $\{\{t \rightarrow b, b \rightarrow w\}\}$ |
| $(b \wedge f) \rightarrow w$ | $\{\{b \rightarrow w\}, \{f \rightarrow w\}\}$ |

*Example 2.5.* Table 1 gives some examples of classical justification for an example knowledge base $\mathcal{K}$ containing the following statements:

(1) Tweety is a bird ($t \rightarrow b$).
(2) Birds fly ($b \rightarrow f$).
(3) If it flies, it has wings ($f \rightarrow w$).
(4) Birds have wings ($b \rightarrow w$).

## 2.3 KLM-Style Defeasible Propositional Logic

The KLM approach involves extending classical logic so as to introduce defeasibility. Ignoring the details for now, the overall idea for propositional logic is to define defeasible analogues of classical entailment $\models$ and classical implication $\rightarrow$. These are denoted $\approx$ and $\mathrel{|\!\sim}$ respectively, so that $\mathcal{K} \approx A \mathrel{|\!\sim} B$ reads as "$\mathcal{K}$ defeasibly entails that $A$ typically implies $B$".[2] The intention then is that unlike classical entailment, defeasible entailment should respect *specificity*. For our penguin example, we should identify that penguins are a specific kind of bird, and that we may need to disregard some statements that apply for birds generally when reasoning about penguins specifically. The following example illustrates this idea using the proper syntax:

*Example 2.6.* Suppose we have a defeasible knowledge base $\mathcal{K}$ of the following statements:

(1) Birds typically fly ($b \mathrel{|\!\sim} f$).
(2) Penguins are typically birds ($p \mathrel{|\!\sim} b$).
(3) Penguins typically do not fly ($p \mathrel{|\!\sim} \neg f$).
(4) King penguins are typically penguins ($k \mathrel{|\!\sim} p$).

We have not yet defined defeasible entailment $\approx$ (and there are many ways to do so) but any sensible definition will reason as if $p$ and $k$ are specific instances of $b$. This means that we can conclude that king penguins and penguins do not fly ($\mathcal{K} \approx k \mathrel{|\!\sim} \neg f$ and $\mathcal{K} \approx p \mathrel{|\!\sim} \neg f$) even though they are birds ($\mathcal{K} \approx k \mathrel{|\!\sim} b$ and $\mathcal{K} \approx p \mathrel{|\!\sim} b$) and birds fly ($\mathcal{K} \approx b \mathrel{|\!\sim} f$). This is in contrast to the classical case where the corresponding classical knowledge base would entail that king penguins and penguins do not exist ($\mathcal{K}' \models \neg k$ and $\mathcal{K}' \models \neg p$ where $\mathcal{K}'$ is the corresponding classical knowledge base).

Having discussed the general intuition, we now present these ideas formally:

*Definition 2.7.* A *statement* is an expression of the form $A \mathrel{|\!\sim} B$ where $A$ and $B$ are propositional formulas ($A \in \mathcal{L}, B \in \mathcal{L}$). A finite set of statements is a (defeasible) knowledge base. From here on we will assume that knowledge bases are defeasible; classical knowledge bases will be designated as such.

[2]When Kraus, Lehmann and Magidor [14] initially set out KLM they described defeasibility with $\mathrel{|\!\sim}$ denoting a consequence relation; however, the approach presented here is now most commonly used [7, 13].

Note that this implies that defeasible implication $\mid\sim$ is the outermost operation in every statement and certainly $\mid\sim$ cannot be nested. This is unlike classical implication $\rightarrow$ for propositional formulas where formulas such as $(A \rightarrow B) \rightarrow C$ are allowed.

Unlike the classical case, there is more than one way to define defeasible entailment and different definitions correspond to different styles of reasoning. As we proceed with our discussion of KLM-style defeasibility, we will introduce concrete notions of defeasible entailment such as *Rational Closure* [17] and *Relevant Closure* [4]. For now, the following definition captures defeasible entailment abstractly:

*Definition 2.8.* A defeasible entailment relation $\approx$ is a binary relation over knowledge bases and statements. For a knowledge base $\mathcal{K}$ and a statement $A \mid\sim B$,

$$\mathcal{K} \approx A \mid\sim B \text{ if } (\mathcal{K}, A \mid\sim B) \in \approx.$$

Lehmann and Magidor [17] propose a series of postulates that define *rational* defeasible entailment, where each postulate can be thought of as asserting an intuitive characteristic we expect of sensible entailment relations (hence the name *rational*). In addition to this axiomatic definition, rational entailment relations have a model-theoretic semantics [7, 17] which we do not discuss here but which is (in some cases) described exactly by reasoning algorithms of reasonable computational complexity. These reasoning algorithms are a central focus in this paper.

Rational Closure is the form of defeasible entailment initially proposed by Lehmann and Magidor [17] and represents a particularly conservative style of reasoning. Rational Closure is rational and therefore characterised both by the postulates for rationality and by a model-theoretic semantics. A significant advantage of Rational Closure is that the reasoning algorithm here is particularly well-behaved in terms of computational complexity. However, the fact that Rational Closure is so conservative (i.e. inferentially weak) can be seen as a shortcoming.

Casini et al. [4] derive Relevant Closure as an adapatation of the reasoning algorithm for Rational Closure with the goal of improving inferential strength while still controlling the computational complexity of the resulting algorithm. Relevant Closure seems to meet these objectives to some extent but unfortunately is not rational which can be seen as a weakness [4].

In the following sections, we define Rational and Relevant Closure in terms of their reasoning algorithms as this is the most convenient perspective from which to approach defeasible explanation. We also note as an aside that a third form of defeasible entailment for KLM by the name of *Lexicographic Closure* has been described in the literature [16] which we do not consider in our analysis.

## 2.4 Rational Closure

Before defining Rational Closure entailment $\approx_{\mathrm{RC}}$, we first need to introduce some preliminary ideas:

*Definition 2.9.* The *materialisation* or *material counterpart* of a knowledge base $\mathcal{K}$, denoted $\overline{\mathcal{K}}$, is the classical knowledge base

$$\{A \rightarrow B \mid A \mid\sim B \in \mathcal{K}\}.$$

In other words, the materialisation of a knowledge base $\mathcal{K}$ is the classical knowledge base containing a corresponding classical implication $A \rightarrow B$ for each statement $A \mid\sim B$ in $\mathcal{K}$. Later we will see that it is also useful to apply this transformation in reverse which motivates the following definition:

*Definition 2.10.* The *defeasible counterpart* of a classical knowledge base $\mathcal{K}$, denoted $\underline{\mathcal{K}}$, where every formula in $\mathcal{K}$ is of the form $A \rightarrow B$ with $A \in \mathcal{L}, B \in \mathcal{L}$ is the knowledge base

$$\{A \mid\sim B \mid A \rightarrow B \in \mathcal{K}\}.$$

Applying the idea of knowledge base materialisation, we can define the *exceptionality* of a propositional formula with respect to a knowledge base:

*Definition 2.11.* A propositional formula $A$ (i.e. $A \in \mathcal{L}$) is exceptional for a knowledge base $\mathcal{K}$ if $\overline{\mathcal{K}} \models \neg A$.

*Example 2.12.* Suppose $\mathcal{K} = \{b \mid\sim f, p \mid\sim b, p \mid\sim \neg f\}$. Then $p$ is exceptional for $\mathcal{K}$ while $b$ is not. This illustrates the intuitive meaning of exceptionality: we can only reason about implications of exceptional formulas by disregarding the more general statements in the knowledge base (refer back to our discussion in Example 2.6).

A related notion is the function $\varepsilon$ which gives us the statements that have an exceptional antecedent in a knowledge base:

*Definition 2.13.* For a knowledge base $\mathcal{K}$, let $\varepsilon(\mathcal{K})$ be the knowledge base

$$\{A \mid\sim B \mid A \mid\sim B \in \mathcal{K} \text{ with } \overline{\mathcal{K}} \models \neg A\}.$$

We are now in a position to define a sequence of knowledge bases $\mathcal{E}_0^{\mathcal{K}}, \mathcal{E}_1^{\mathcal{K}}, \cdots, \mathcal{E}_n^{\mathcal{K}}$ such that knowledge bases earlier in the sequence contain, in addition to the statements in later knowledge bases, statements that are more defeasible (i.e. less specific) than those in later knowledge bases [13]:

*Definition 2.14.* For a knowledge base $\mathcal{K}$, the *exceptionality sequence* $\mathcal{E}$ is given by letting $\mathcal{E}_0^{\mathcal{K}} = \mathcal{K}$ and $\mathcal{E}_{i+1}^{\mathcal{K}} = \varepsilon(\mathcal{E}_i^{\mathcal{K}})$ for $1 \le i \le n$. The index $n$ of the final knowledge base is the smallest $i$ such that $\varepsilon(\mathcal{E}_i^{\mathcal{K}}) = \mathcal{E}_i^{\mathcal{K}}$.

The final element is often identified with infinity so that $\mathcal{E}_\infty^{\mathcal{K}}$ is the same as $\mathcal{E}_n^{\mathcal{K}}$. This final knowledge base $\mathcal{E}_\infty^{\mathcal{K}}$ is unique in that its statements are not retractable[3] (i.e. cannot be disregarded) and note that we may have $\mathcal{E}_\infty^{\mathcal{K}} = \emptyset$ in the event that $\mathcal{K}$ does not contain any non-retractable statements [7, 13].

One of the consequences of how we define $\mathcal{E}$ is that any statement of the form $\neg A \mid\sim \bot$ in $\mathcal{K}$ where $A$ is a propositional formula will have $\neg A \mid\sim \bot \in \mathcal{E}_\infty^{\mathcal{K}}$. Since the statements in $\mathcal{E}_\infty^{\mathcal{K}}$ are not retractable, this allows us to express classical (i.e. categorical) information in a defeasible knowledge base: $\neg A \mid\sim \bot$ encodes the classical assertion made in $A$ in a defeasible knowledge base. It is often useful to express classical formulas in a defeasible context in this manner, but for the sake of compactness, we will often write $A$ in favour of $\neg A \mid\sim \bot$. This is a shorthand and our meaning is always the latter.

We also define a ranking $\mathcal{R}_0^{\mathcal{K}}, \cdots, \mathcal{R}_\infty^{\mathcal{K}}$ of the statements in a knowledge base $\mathcal{K}$ which follows the same principle but has that each rank is disjoint from the others:

[3] This is apparent when we define Rational Closure entailment $\approx_{\mathrm{RC}}$ in Definition 2.19.

| | |
|---|---|
| $\infty$ | $p \rightarrow b, r \rightarrow p$ |
| 1 | $p \mathrel{\vert\!\sim} \neg f$ |
| 0 | $b \mathrel{\vert\!\sim} f, b \mathrel{\vert\!\sim} w$ |

**Figure 1: Ranking $\mathcal{R}_0^{\mathcal{K}}, \cdots, \mathcal{R}_\infty^{\mathcal{K}}$ for Example 2.16**

*Definition 2.15.* For a knowledge base $\mathcal{K}$, let $\mathcal{R}_i^{\mathcal{K}} = \mathcal{E}_i^{\mathcal{K}} \setminus \mathcal{E}_{i+1}^{\mathcal{K}}$ for $0 \leq i \leq n-1$ and $\mathcal{R}_\infty^{\mathcal{K}} = \mathcal{E}_\infty^{\mathcal{K}}$.

*Example 2.16.* Consider the knowledge base $\mathcal{K}$ containing the following statements:

(1) Birds typically fly ($b \mathrel{\vert\!\sim} f$).
(2) Birds typically have wings ($b \mathrel{\vert\!\sim} w$).
(3) Penguins are birds ($p \rightarrow b$).
(4) Penguins typically do not fly ($p \mathrel{\vert\!\sim} \neg f$).
(5) Rico is a penguin ($r \rightarrow p$).

The ranking $\mathcal{R}_0^{\mathcal{K}}, \cdots, \mathcal{R}_\infty^{\mathcal{K}}$ for $\mathcal{K}$ is given in Figure 1. Note that as we suggested above, $p \rightarrow b$ and $r \rightarrow p$ are shorthands for $\neg(p \rightarrow b) \mathrel{\vert\!\sim} \bot$ and $\neg(r \rightarrow p) \mathrel{\vert\!\sim} \bot$ respectively. We find that rules about more specific objects ($p$) appear higher than rules about less specific objects ($b$) and classical information appears in $\mathcal{R}_\infty^{\mathcal{K}}$.

We now introduce *base ranks*, a concept that is central to our definition of defeasible entailment:

*Definition 2.17.* The base rank $\mathrm{br}^{\mathcal{K}}(A)$ of a propositional formula $A$ is the smallest index $i$ such that $A$ is not exceptional for $\mathcal{E}_i^{\mathcal{K}}$:

$$\mathrm{br}^{\mathcal{K}}(A) = \min\left\{ i \mid \overline{\mathcal{E}_i^{\mathcal{K}}} \not\models \neg A \right\}.$$

If there is no $i$ for which $\overline{\mathcal{E}_i^{\mathcal{K}}} \not\models \neg A$, then let $\mathrm{br}^{\mathcal{K}}(A) = \infty$. This is distinguished from the case of $\mathrm{br}^{\mathcal{K}}(A) = n$ (where $\mathcal{E}_\infty^{\mathcal{K}}$ is the first $\mathcal{E}_i^{\mathcal{K}}$ such that $\overline{\mathcal{E}_i^{\mathcal{K}}} \not\models \neg A$).

We also introduce the following shorthand:

*Definition 2.18.* For a knowledge base $\mathcal{K}$ and a propositional formula $A$, let $\mathcal{E}_A^{\mathcal{K}} = \mathcal{E}_r^{\mathcal{K}}$ where $r = \mathrm{br}^{\mathcal{K}}(A)$. The cases of $r = \infty$ and $r = n$ both correspond to $\mathcal{E}_A^{\mathcal{K}} = \mathcal{E}_\infty^{\mathcal{K}}$.

We can now define Rational Closure entailment [11, 13, 17]:

*Definition 2.19.* For a knowledge base $\mathcal{K}$ and an entailment query $A \mathrel{\vert\!\sim} B$ ($A \in \mathcal{L}, B \in \mathcal{L}$),

$$\mathcal{K} \approx_{\mathrm{RC}} A \mathrel{\vert\!\sim} B \text{ if } \mathrm{br}^{\mathcal{K}}(A) = \infty \text{ or } \mathrm{br}^{\mathcal{K}}(A) < \mathrm{br}^{\mathcal{K}}(A \land \neg B).$$

Since we never have $\mathrm{br}^{\mathcal{K}}(A) > \mathrm{br}^{\mathcal{K}}(A \land \neg B)$, we can test for $\approx_{\mathrm{RC}}$ entailment simply by evaluating whether $A \land \neg B$ is exceptional for $\mathcal{E}_A^{\mathcal{K}}$:

Proposition 2.20. *For a knowledge base $\mathcal{K}$ and an entailment query $A \mathrel{\vert\!\sim} B$,*

$$\mathcal{K} \approx_{\mathrm{RC}} A \mathrel{\vert\!\sim} B \textit{ iff } \mathrm{br}^{\mathcal{K}}(A) = \infty \textit{ or } \overline{\mathcal{E}_A^{\mathcal{K}}} \models A \rightarrow B.$$

Note that strictly speaking we do not need $\mathrm{br}^{\mathcal{K}}(A) = \infty$ as a special case here because the latter condition is always true for $\mathrm{br}^{\mathcal{K}}(A) = \infty$.

**Table 2: Examples of $\approx_{\mathrm{RC}}$ entailment for Example 2.21**

| $A \mathrel{\vert\!\sim} B$ | $\mathrm{br}^{\mathcal{K}}(A)$ | $\overline{\mathcal{E}_A^{\mathcal{K}}} \models A \rightarrow B$? | Result |
|---|---|---|---|
| $b \mathrel{\vert\!\sim} \neg f$ | 0 | No | $\mathcal{K} \not\approx_{\mathrm{RC}} b \mathrel{\vert\!\sim} \neg f$ |
| $r \mathrel{\vert\!\sim} \neg f$ | 1 | Yes | $\mathcal{K} \approx_{\mathrm{RC}} r \mathrel{\vert\!\sim} \neg f$ |
| $p \mathrel{\vert\!\sim} w$ | 1 | No | $\mathcal{K} \not\approx_{\mathrm{RC}} p \mathrel{\vert\!\sim} w$ |

*Example 2.21.* Consider the knowledge base $\mathcal{K}$ from Example 2.16:

$$\mathcal{K} = \{ b \mathrel{\vert\!\sim} f, b \mathrel{\vert\!\sim} w, p \rightarrow b, p \mathrel{\vert\!\sim} \neg f, r \rightarrow p \}.$$

We gave $\mathcal{R}_0^{\mathcal{K}}, \cdots, \mathcal{R}_\infty^{\mathcal{K}}$ in Figure 1. We evaluate some sample entailment queries for $\mathcal{K}$ in Table 2 by applying Proposition 2.20. Intuitively, the idea here is that we start with $\mathcal{K}$ and keep removing the lowest rank:

$$\mathcal{E}_0^{\mathcal{K}} = \mathcal{K}; \mathcal{E}_1^{\mathcal{K}} = \mathcal{K} \setminus \mathcal{R}_0^{\mathcal{K}}; \mathcal{E}_2^{\mathcal{K}} = \mathcal{K} \setminus (\mathcal{R}_0^{\mathcal{K}} \cup \mathcal{R}_1^{\mathcal{K}}); \ldots$$

We repeat this until we find the first $\mathcal{E}_i^{\mathcal{K}}$ for which the antecedent of the query is not exceptional (or otherwise $\mathcal{E}_\infty^{\mathcal{K}}$ if there is none). Then, having eliminated the more general statements in lower ranks, we use classical entailment to reason about the knowledge base:

$$\mathcal{K} \approx_{\mathrm{RC}} A \mathrel{\vert\!\sim} B \text{ iff } \overline{\mathcal{E}_i^{\mathcal{K}}} \models A \rightarrow B.$$

Apart from illustrating reasoning using Rational Closure, the example above shows why Rational Closure is inferentially weak: notice how we do not conclude that penguins have wings ($\mathcal{K} \not\approx_{\mathrm{RC}} p \mathrel{\vert\!\sim} w$) even though penguins are birds and birds have wings ($p \mathrel{\vert\!\sim} b, b \mathrel{\vert\!\sim} w$). In other words, more specific formulas never inherit the implications of less specific formulas even when doing so would not involve a contradiction. This is not a problem here—penguins, after all, do not have wings—but this may be undesirable in other circumstances. In terms of the reasoning algorithm, the reason for this is that Rational Closure retracts entire ranks of less specific statements even though only a handful may be responsible for the exceptionality of the antecedent.

Algorithms 1 and 2 express this definition of Rational Closure procedurally. The former computes the sequence $\mathcal{R}_0^{\mathcal{K}}, \cdots, \mathcal{R}_\infty^{\mathcal{K}}$ for a knowledge base $\mathcal{K}$ and the latter computes whether $\mathcal{K} \approx_{\mathrm{RC}} A \mathrel{\vert\!\sim} B$ for a knowledge base $\mathcal{K}$ and a query $A \mathrel{\vert\!\sim} B$ [10, 17]. Entailment checking for Rational Closure is computationally well-behaved since the number of classical entailments we need to evaluate in these algorithms is a polynomial function of the size of the knowledge base [13].

## 2.5 Relevant Closure

We mentioned that a problem of Rational Closure is that it is too conservative, and that the reason for this is that it retracts more information than is intuitively necessary. As a solution to this problem, Casini et. al. [4] propose Relevant Closure which adapts Rational Closure so that we only retract the statements in a less specific rank that actually disagree with more specific statements in higher ranks with respect to the antecedent of the query. Casini et al. in fact describe two forms of Relevant Closure, *Basic Relevant Closure* and *Minimal Relevant Closure*, where the former is more conservative than the latter. Relevant Closure was initially presented in terms of the KLM description logic but the same ideas

**ALGORITHM 1:** Rank

**Data:** A knowledge base $\mathcal{K}$
**Result:** $\left(\mathcal{R}_0^{\mathcal{K}}, \cdots, \mathcal{R}_n^{\mathcal{K}}, n\right)$

$i := 0$
$E_0 := \mathcal{K}$
**while** $E_i \neq \varepsilon(E_i)$ **do**
  $E_{i+1} := \varepsilon(E_i)$
  $R_i := E_i \setminus E_{i+1}$
  $i := i + 1$
**end**
$R_i := E_i$
**return** $(R_0, \cdots, R_i, i)$

**ALGORITHM 2:** RationalClosure

**Data:** A knowledge base $\mathcal{K}$ and a query $A \mid\sim B$
**Result:** TRUE iff $\mathcal{K} \mathrel{|\!\approx}_{\mathrm{RC}} A \mid\sim B$

$(\mathcal{K}_0, \cdots, \mathcal{K}_n, n) :=$ Rank($\mathcal{K}$)
$i := 0$
$\mathcal{K}' := \mathcal{K}$
**while** $i < n$ *and* $\overline{\mathcal{K}'} \models \neg A$ **do**
  $\mathcal{K}' := \mathcal{K}' \setminus \mathcal{K}_i$
  $i := i + 1$
**end**
**return** $\overline{\mathcal{K}'} \models A \to B$

are applicable to KLM propositional logic and we give a definition in these terms. We begin by defining a notion of justification distinct from, but very much related to, the classical justifications we discussed in Section 2.2:

*Definition 2.22.* For a knowledge base $\mathcal{K}$ and $A \in \mathcal{L}$, $\mathcal{J}_\varepsilon$ is an *ε-justifcation* for the pair $(\mathcal{K}, A)$ if $\mathcal{J}_\varepsilon$ is a minimal subset $\mathcal{J}_\varepsilon \subseteq \mathcal{K}$ such that $A$ is exceptional for $\mathcal{J}_\varepsilon$. By minimal we mean that there is no proper subset $\mathcal{J}_\varepsilon{}' \subset \mathcal{J}_\varepsilon$ such that $A$ is exceptional for $\mathcal{J}_\varepsilon{}'$. Denote the set of ε-justifications for $(\mathcal{K}, A)$ as $\mathcal{J}_\varepsilon(\mathcal{K}, A)$.

This notion of justification is closely connected to classical justification:

Proposition 2.23. *A knowledge base $\mathcal{J}_\varepsilon$ is an ε-justification for $(\mathcal{K}, A)$ iff $\overline{\mathcal{J}_\varepsilon}$ is a classical justification for $\overline{\mathcal{K}} \models \neg A$.*

This concept allows us to introduce *basic* and *minimal relevance*:

*Definition 2.24.* Given a knowledge base $\mathcal{K}$ and a query $A \mid\sim B$, the statements basically relevant to $A \mid\sim B$ are those that appear in an ε-justification for $(\mathcal{K}, A)$, i.e. $\bigcup \mathcal{J}_\varepsilon(\mathcal{K}, A)$. On the other hand, the statements minimally relevant to $A \mid\sim B$ are

$$\bigcup \{\min_{\mathrm{br}} \mathcal{J}_\varepsilon \mid \mathcal{J}_\varepsilon \in \mathcal{J}_\varepsilon(\mathcal{K}, A)\}$$

where $\min_{\mathrm{br}} \mathcal{J}_\varepsilon$ gives the statements in $\mathcal{J}_\varepsilon$ that have antecedents with the smallest base rank:

$$\min_{\mathrm{br}} \mathcal{J}_\varepsilon = \Big\{C \mid\sim D \mid C \mid\sim D \in \mathcal{J}_\varepsilon, \text{ and for all } E \mid\sim F \in \mathcal{J}_\varepsilon, \mathrm{br}^{\mathcal{K}}(C) \leq \mathrm{br}^{\mathcal{K}}(E)\Big\}.$$

Given some query $A \mid\sim B$, relevance (either minimal or basic) enables us to *partition* a knowledge base $\mathcal{K}$ into a part that is relevant to the query and a part that is not relevant to the query:

*Definition 2.25.* For a notion of relevance (basic or minimal) and a query $A \mid\sim B$, the relevance partition $(R, R^-)$ of a knowledge base $\mathcal{K}$ is given by

$$R = \{C \mid\sim D \mid C \mid\sim D \in \mathcal{K} \text{ is relevant to } A \mid\sim B\},$$
$$R^- = \mathcal{K} \setminus R.$$

We then define Relevant Closure entailment in a way that strongly resembles reasoning for Rational Closure (Proposition 2.20), only in this case we never retract statements that are considered not relevant to the query:

*Definition 2.26.* Given a relevance partition $(R, R^-)$ for a knowledge base $\mathcal{K}$ and a query $A \mid\sim B$, both forms of Relevant Closure are defined as:

$$\mathcal{K} \mathrel{|\!\approx} A \mid\sim B \text{ if } \mathrm{br}^{\mathcal{K}}(A) = \infty \text{ or } \overline{\mathcal{E}_A^{\mathcal{K}} \cup R^-} \models A \to B.$$

Basic Relevant Closure entailment $\mathrel{|\!\approx}_{\mathrm{BRelC}}$ uses the basic relevance partition and Minimal Relevant Closure $\mathrel{|\!\approx}_{\mathrm{MRelC}}$ uses the minimal relevance partition.

*Example 2.27.* Consider the knowledge base $\mathcal{K}$ from Example 2.16:

$$\mathcal{K} = \{b \mid\sim f, b \mid\sim w, p \to b, p \mid\sim \neg f, r \to p\}.$$

Suppose we wish to test if Rico typically has wings ($\mathcal{K} \mathrel{|\!\approx} r \mid\sim w$) for Basic or Minimal Relevant Closure. We gave $\mathcal{R}_0^{\mathcal{K}}, \cdots, \mathcal{R}_\infty^{\mathcal{K}}$ in Figure 1. We have $\mathrm{br}^{\mathcal{K}}(r) = 1$ and hence

$$\mathcal{E}_r^{\mathcal{K}} = \{p \mid\sim \neg f, p \to b, r \to p\}.$$

However, we do not retract the entirety of $\mathcal{R}_0^{\mathcal{K}}$ since $b \mid\sim w$ is not relevant for the query $r \mid\sim w$ (for either basic or minimal relevance) and therefore remains under our consideration:

$$\mathcal{E}_r^{\mathcal{K}} \cup R^- = \{p \mid\sim \neg f, p \to b, r \to p, b \mid\sim w\}$$

for which we have $\overline{\mathcal{E}_A^{\mathcal{K}} \cup R^-} \models r \to w$ and thus $\mathcal{K} \mathrel{|\!\approx}_{\mathrm{BRelC}} r \mid\sim w$ and $\mathcal{K} \mathrel{|\!\approx}_{\mathrm{MRelC}} r \mid\sim w$.

The example above illustrates that Relevant Closure, unlike Rational Closure, allows specific formulas to inherit implications of less specific formulas so long as they are not contradictory. This is why Relevant Closure is inferentially stronger than Rational Closure. We summarise the reasoning algorithm for Relevant Closure in Algorithm 3; it is derived as an adaptation of Algorithm 2. The computational complexity of this algorithm is greater than that for Rational Closure since we now need to enumerate justifications although the difference is not unreasonable [4].

**ALGORITHM 3:** RelevantClosure

**Data:** A knowledge base $\mathcal{K}$, query $A \mid\sim B$ and a partition $(R, R^-)$
**Result:** TRUE iff $\mathcal{K} \mathrel{|\!\approx} A \mid\sim B$ for the form of Relevant Closure at hand

$(\mathcal{K}_0, \cdots, \mathcal{K}_n, n) :=$ Rank($\mathcal{K}$)
$i := 0$
$\mathcal{K}' := \mathcal{K}$
**while** $i < n$ *and* $\overline{\mathcal{K}'} \models \neg A$ **do**
  $\mathcal{K}' := \mathcal{K}' \setminus (\mathcal{K}_i \cap R)$
  $i := i + 1$
**end**
**return** $\overline{\mathcal{K}'} \models A \to B$

| | |
|---|---|
| $\infty$ | |
| 2 | $s \mid\sim p, s \mid\sim f$ |
| 1 | $p \mid\sim b, p \mid\sim \neg f$ |
| 0 | $b \mid\sim f$ |

**Figure 2: Ranking $\mathcal{R}_0^{\mathcal{K}}, \cdots, \mathcal{R}_\infty^{\mathcal{K}}$ for Example 3.1**

# 3 DEFEASIBLE EXPLANATION

Unlike the classical case, explanation has not yet been explored in detail for defeasible reasoning [13] apart from some introductory work [3, 9]. Before discussing this work, it will be instructive to consider why a naive definition of defeasible justification is insufficient. In particular, it may be tempting to simply define defeasible justification similarly to how we define classical justification; namely, to say that $\mathcal{J}$ is a justification for $\mathcal{K} \approx A$ if $\mathcal{J}$ is a minimal subset such that $\mathcal{J} \approx A$. Unfortunately, this is not a sensible definition. The following example illustrates the issue here:

*Example 3.1.* Consider the knowledge base $\mathcal{K}$ containing the following statements:

(1) Birds typically fly ($b \mid\sim f$).
(2) Penguins are typically birds ($p \mid\sim b$).
(3) Penguins typically do not fly ($p \mid\sim \neg f$).
(4) Special penguins are typically penguins ($s \mid\sim p$)
(5) Special penguins typically fly ($s \mid\sim f$).

The ranking for $\mathcal{K}$ is given in Figure 2. Suppose we are justifying the entailment that special penguins typically fly with Rational Closure ($\mathcal{K} \approx_{\text{RC}} s \mid\sim f$) using the naive definition above. The minimal subsets of $\mathcal{K}$ that defeasibly entail $s \mid\sim f$ are $\{s \mid\sim p, s \mid\sim f\}$ and $\{s \mid\sim p, p \mid\sim b, b \mid\sim f\}$. We might intuitively think that the former ought to be a justification, but the latter is problematic: we are saying that special penguins fly not because of their specific rule ($s \mid\sim f$) but because special penguins are typically birds and birds typically fly ($s \mid\sim p, p \mid\sim b, b \mid\sim f$). Apart from being intuitively incorrect, this does not at all correspond to the reasoning formalism: the statements $p \mid\sim b$ and $b \mid\sim f$ are retracted as we find that more specific statements ($s \mid\sim p, s \mid\sim f$) disagree with respect to the the antecedent $s$ of the query.

## 3.1 Weak Explanation

One of the main works of interest here is that of Chama [9] which proposes a notion of defeasible entailment for Rational Closure according to an algorithm closely connected to the Rational Closure reasoning algorithm (Algorithm 2). The insight here is that we should follow the same process to eliminate more general statements, and once we have done so, to use classical tools to reason about the knowledge base—only in this case we obtain classical justifications instead of testing for a classical entailment. We refer to these justifications as *weak justifications* to distinguish from classical justifications and the strong justifications we discuss later. Though initially given in terms of KLM description logic, we apply this result to KLM propositional logic in Definition 3.2 and Algorithm 4:

*Definition 3.2.* A knowledge base $\mathcal{J}$ is a weak justification for an entailment $\mathcal{K} \approx_{\text{RC}} A \mid\sim B$ if $\mathcal{J}$ is an element of the set returned by Algorithm 4 given $\mathcal{K}$ and $A \mid\sim B$. This set of justifications returned by Algorithm 4 is denoted $\mathcal{J}_W(\mathcal{K}, A \mid\sim B)$.

**ALGORITHM 4:** `WeakJustifyRC`

**Data:** A knowledge base $\mathcal{K}$ and a query $A \mid\sim B$
**Result:** The set of weak justifications for $\mathcal{K} \approx_{\text{RC}} A \mid\sim B$

$(\mathcal{K}_0, \cdots, \mathcal{K}_n, n) := \mathsf{Rank}(\mathcal{K})$
$i := 0$
$\mathcal{K}' := \mathcal{K}$
**while** $i < n$ *and* $\overline{\mathcal{K}'} \models \neg A$ **do**
  $\mathcal{K}' := \mathcal{K}' \setminus \mathcal{K}_i$
  $i := i + 1$
**end**
**return** $\left\{\underline{\mathcal{J}} \mid \mathcal{J} \in \mathcal{J}\left(\overline{\mathcal{K}'}, A \to B\right)\right\}$

As an alternative to this procedural definition, it can be shown that Definition 3.2 is equivalent to the following:

Proposition 3.3. *A knowledge base $\mathcal{J}$ is a weak justification for an entailment $\mathcal{K} \approx_{\text{RC}} A \mid\sim B$ iff $\overline{\mathcal{J}}$ is a classical justification for $\overline{\mathcal{E}_A^{\mathcal{K}}} \models A \to B$.*

*Example 3.4.* Consider the knowledge base $\mathcal{K}$ and the query $s \mid\sim f$ from Example 3.1:

$$\mathcal{K} = \{b \mid\sim f, p \mid\sim b, p \mid\sim \neg f, s \mid\sim p, s \mid\sim f\}\,.$$

The ranking $\mathcal{R}_0^{\mathcal{K}}, \cdots, \mathcal{R}_\infty^{\mathcal{K}}$ is given in Figure 2. We have $\mathcal{E}_s^{\mathcal{K}} = \{s \mid\sim p, s \mid\sim f\}$. Then we look at the classical justifications for

$$\overline{\mathcal{E}_s^{\mathcal{K}}} \models s \to f$$

and clearly $\{s \to f\}$ is the sole classical justification here, so by Proposition 3.3 $\{s \mid\sim f\}$ is a unique weak justification for $\mathcal{K} \approx_{\text{RC}} A \mid\sim B$.

The results weak justification produces are intuitive and the correspondence between weak justification and the notion of defeasible entailment $\approx_{\text{RC}}$ is unmistakable. We also have a clear understanding of how to compute these justifications and the algorithm is computationally well-behaved [9]. As presented, however, weak justification is limited to the case of Rational Closure which may limit its utility given that Rational Closure is inferentially weak. In addition, weak justification has been presented purely in terms of the reasoning algorithms (such constructions as the exceptionality sequence $\mathcal{E}$ and base ranks) and therefore an interesting question is whether weak justification can be characterised in a more intuitive manner. We will spend much of this paper addressing these concerns. Before presenting these results, we consider a different approach to defeasible explanation known as *strong explanation*.

## 3.2 Strong Explanation

Brewka and Ulbricht [3] suggest a different approach to defeasible explanation by imposing an additional constraint on the naive definition which requires that additional information from $\mathcal{K}$ cannot be added to the justification so as to render the entailment false. Brewka and Ulbricht consider this result both abstractly and in terms of logic programs but the principle is easily applied to KLM defeasible propositional logic:

*Definition 3.5.* A knowledge base $\mathcal{J}$ is a *strong justification* for an entailment $\mathcal{K} \approx A \vdash B$ if $\mathcal{J}$ is a minimal subset $\mathcal{J} \subseteq \mathcal{K}$ such that $\mathcal{J} \approx A \vdash B$ and there is no superset $\mathcal{J}' \supseteq \mathcal{J}$ where $\mathcal{J}' \subseteq \mathcal{K}$ such that $\mathcal{J}' \not\approx A \vdash B$. The subset is minimal if it has no proper subset that satisfies these conditions.

Note that we have not restricted ourselves to a particular formalism of defeasible entailment such as $\approx_{\mathrm{RC}}$ or $\approx_{\mathrm{MRelC}}$. Indeed, one of the apparent strengths of strong justification is that its definition is abstract and declarative enough to potentially be widely applicable not only to different entailment formalisms but also to many different frameworks for defeasible reasoning [3]. The definition also seems to express an idea that is simple and perhaps intuitive, namely, that justifications should be such that they always defeasibly entail the query even when they are combined with arbitrary portions of the knowledge base. This is in contrast to our discussion of weak justification which so far has been constructive rather than axiomatic. With that in mind, we now explore how weak and strong justifications relate to each other and whether this is a sound notion of justification as far as KLM is concerned, starting with a simple example:

*Example 3.6.* Consider the entailment $\mathcal{K} \approx_{\mathrm{RC}} s \vdash f$ from Example 3.1. We found in that example that the minimal subsets of $\mathcal{K}$ that defeasibly entail $s \vdash f$ are

$$\mathcal{J}_1 = \{s \vdash f\}, \mathcal{J}_2 = \{s \vdash p, p \vdash b, b \vdash f\}.$$

We identify that $\mathcal{J}_2$ is not a strong justification because for example

$$\mathcal{J}_2 \cup \{p \vdash \neg f\} \not\approx_{\mathrm{RC}} s \vdash f.$$

On the other hand, $\mathcal{J}_1$ is a strong justification because there is no superset $\mathcal{J}'$ of $\mathcal{J}_1$ such that $\mathcal{J}' \not\approx_{\mathrm{RC}} A \vdash B$.[4]

Since we had the same results in Examples 3.4 and 3.6, a natural question at this point is to ask whether weak and strong justifications describe the same underlying concept, or if one condition is sufficient for the other, i.e. if

$$\mathcal{J} \text{ is weak justification} \implies \mathcal{J} \text{ is a strong justification}$$

or vice versa for a given $\approx_{\mathrm{RC}}$ entailment. The answer is that there is no such connection. The following example illustrates the distinction between the two concepts:

*Example 3.7.* Consider $\mathcal{K} \approx_{\mathrm{RC}} s \wedge p \vdash f$ for the knowledge base $\mathcal{K}$ from Example 3.1:

$$\mathcal{K} = \{b \vdash f, p \vdash b, p \vdash \neg f, s \vdash p, s \vdash f\}.$$

Using the ranking in Figure 3.1 we find that $\mathcal{E}^{\mathcal{K}}_{s \wedge p} = \{s \vdash p, s \vdash f\}$ and hence that $\mathcal{J} = \{s \vdash f\}$ is the only weak justification for the entailment. However, $\mathcal{J}$ is not a strong justification since

$$\mathcal{J}' = \{s \vdash f, p \vdash \neg f\} \text{ has } \mathcal{J}' \not\approx_{\mathrm{RC}} s \wedge p \vdash f.$$

The reason for this is that the statement $s \vdash p$ ensures that $s \vdash f$ is regarded as more specific than $p \vdash \neg f$ and when it is omitted $s \vdash f$ and $p \vdash \neg f$ are both placed in the zeroth rank $\mathcal{R}^{\mathcal{J}'}_0$. Then since

$$\{s \to f, p \to \neg f\} \models \neg(s \wedge p)$$

we have $\mathcal{E}^{\mathcal{J}'}_{s \wedge p} = \emptyset$ and the defeasible entailment does not hold. In fact, we obtain the (unique) strong justification here by adding $s \vdash p$ to $\mathcal{J}$ to give us $\mathcal{J}_S = \{s \vdash p, s \vdash f\}$ for which there is no superset that does not defeasibly entail the query.

This example, apart from illustrating that weak justification and strong justification are different, suggests that strong justification might call for a different intuitive interpretation than weak justification. At least in this example, the strong justification seems to be somewhat more comprehensive than the weak justification: not only do we have $s \vdash f \in \mathcal{J}_S$—the most specific and therefore applicable rule for the query—but we also have $s \vdash p \in \mathcal{J}_S$ which ensures that $s \vdash f$ is actually regarded as the most specific rule. This is an interesting distinction, though we should perhaps not be surprised that there might be different, sensible notions of defeasible justification given the nuances introduced by nonmonotonicity (just as there were several viable notions of defeasible entailment). It is worth pointing out however that this intuition for how strong justifications differ from weak justifications may not be entirely accurate as we illustrate in the following example:

| | |
|---|---|
| $\infty$ | |
| 1 | $p \vdash b, p \vdash v \wedge \neg f$ |
| 0 | $b \vdash v, b \vdash f$ |

**Figure 3: Ranking $\mathcal{R}^{\mathcal{K}}_0, \cdots, \mathcal{R}^{\mathcal{K}}_\infty$ for Example 3.8**

*Example 3.8.* Consider the knowledge base $\mathcal{K}$ of the following statements:

(1) Penguins are typically vertebrates and typically do not fly ($p \vdash v \wedge \neg f$).
(2) Penguins are typically birds ($p \vdash b$).
(3) Birds typically fly ($b \vdash f$).
(4) Birds are typically vertebrates ($b \vdash v$).

Consider the $\approx_{\mathrm{RC}}$ entailment that penguins are typically vertebrates ($\mathcal{K} \approx_{\mathrm{RC}} p \vdash v$). We give $\mathcal{R}^{\mathcal{K}}_0, \cdots, \mathcal{R}^{\mathcal{K}}_\infty$ in Figure 3. We find that

$$\{p \vdash v \wedge \neg f\}$$

is both a weak and strong justification for this entailment. However, we also find that

$$\mathcal{J}_S = \{p \vdash b, b \vdash v\}$$

is a strong justification while there are no other weak justifications. Intuitively, the reason for this is that even though $b \vdash v$ is not a specific rule for the antecedent $p$ given $\mathcal{K}$ (i.e. $b \vdash v \notin \mathcal{E}^{\mathcal{K}}_p$), there happens to be no way to add statements to $\mathcal{J}_S$ so as to render the entailment false. When either $p \vdash v \wedge \neg f$ or $b \vdash f$ is excluded, all statements are in the zeroth rank and $b \vdash v$ is specific for $p$. When $p \vdash v \wedge \neg f, b \vdash f$ are included, statements about $p$ appear above the zeroth rank and $b \vdash v$ is no longer specific for $p$ but then of course $p \vdash v \wedge \neg f$ serves as a specific assertion that penguins are vertebrates. If we had expressed that penguins are flightless vertebrates using two separate statements—$p \vdash v, p \vdash \neg f$ rather than $p \vdash v \wedge \neg f$—then $\mathcal{J}_S$ would not be a strong justification because

$$\mathcal{J}_S \cup \{b \vdash f, p \vdash \neg f\} \not\approx_{\mathrm{RC}} p \vdash v.$$

This example seems to highlight a potential issue with strong justification as applied to KLM-style defeasible entailment since

[4] There is no obvious way to compute this efficiently. Our strategy for these examples is simply to consider every superset of the candidate justification.

the justification $\mathcal{J}_S$ does not seem to communicate the true reason why penguins are vertebrates according to $\approx_{RC}$ entailment (similar to our discussion of the naive definition in Example 3.1). Moreover, the sensitivity to syntax when comparing $p \mathrel{|\!\sim} v, p \mathrel{|\!\sim} \neg f$ to $p \mathrel{|\!\sim} v \wedge \neg f$ is questionable. Although there is no inherent problem with syntactic dependencies—indeed, our notions of KLM-style defeasible entailment are very closely tied to syntax—this particular instance seems intuitively unprincipled. Sensitivities such as this would presumably be difficult to predict or understand for users of the reasoning system.

A potential strategy here is to restrict our consideration only to those strong justifications that are supersets of some weak justification for the entailment. This produces a better result for the case of Example 3.8 where we would have simply have $\{p \mathrel{|\!\sim} v \wedge \neg f\}$ which is intuitive. In general, this revised definition has no obvious problems and corresponds much better to the notion of defeasible entailment. Of course, this adjustment places weak justification as a more elementary notion of defeasible justification and strong justification as a potentially interesting extension of weak justification. In that regard, while strong justifications may be a promising line of further enquiry, this does perhaps suggest that they do not represent the parsimonious and abstract notion of defeasible justification for KLM we alluded to earlier.

Beyond these issues, another hurdle for strong justifications is that there is no obvious way to compute them efficiently: our attempts to find an efficient algorithm for evaluating strong justifications for KLM were to no avail. Ultimately, our impression is that while there may be value in exploring strong justification further for KLM, weak justification is more predictable and likely the closer defeasible analogue to classical justification.

## 4 PROPERTIES FOR DEFEASIBLE EXPLANATION

We have discussed the appeals of weak justification as a formalism for defeasible justification but we also noted that it currently lacks an intuitive characterisation. The result we present in this section is that weak justification obeys several properties much in the same way that rational entailment relations such as Rational Closure obey the postulates of rational defeasible entailment. First, let us consider a strengthening of the postulates for rationality given by Lehmann and Magidor [13, 17]:[5]

*Definition 4.1.* A defeasible entailment relation $\approx$ is rational (or *LM-rational*) if it obeys the following postulates:

(1) Left logical equivalence *(LLE)*. If $\mathcal{K} \approx A \leftrightarrow B$ and $\mathcal{K} \approx A \mathrel{|\!\sim} C$ then $\mathcal{K} \approx B \mathrel{|\!\sim} C$.
(2) Right weakening *(RW)*. If $\mathcal{K} \approx A \rightarrow B$ and $\mathcal{K} \approx C \mathrel{|\!\sim} A$ then $\mathcal{K} \approx C \mathrel{|\!\sim} B$.
(3) *And*. If $\mathcal{K} \approx A \mathrel{|\!\sim} B$ and $\mathcal{K} \approx A \mathrel{|\!\sim} C$ then $\mathcal{K} \approx A \mathrel{|\!\sim} B \wedge C$.
(4) *Or*. If $\mathcal{K} \approx A \mathrel{|\!\sim} C$ and $\mathcal{K} \approx B \mathrel{|\!\sim} C$ then $\mathcal{K} \approx A \vee B \mathrel{|\!\sim} C$.
(5) Reflexivity *(Ref)*. $\mathcal{K} \approx A \mathrel{|\!\sim} A$.
(6) Cautious Monotonicity *(CM)*. If $\mathcal{K} \approx A \mathrel{|\!\sim} C$ and $\mathcal{K} \approx A \mathrel{|\!\sim} B$ then $\mathcal{K} \approx A \wedge B \mathrel{|\!\sim} C$.
(7) Rational Monotonicity *(RM)*. If $\mathcal{K} \approx A \mathrel{|\!\sim} C$ and $\mathcal{K} \not\approx A \mathrel{|\!\sim} \neg B$ then $\mathcal{K} \approx A \wedge B \mathrel{|\!\sim} C$.

Each postulate expresses a behaviour we intuitively expect of defeasible reasoning. For instance, consider *And*: if $A$ typically implies $B$, and $A$ typically implies $C$, then surely $A$ must typically imply $B$ and $C$. There is a corresponding intuitive interpretation for each of these postulates. These postulates have been studied at length in the literature and have proved useful in the theoretical analysis of KLM-style defeasible entailment [4, 6, 13, 14, 17].

Our approach is to consider how defeasible justification applies to each of these axioms. Take for instance the example of *And*. The insight here is that if $A$ typically implies $B$, and $A$ typically implies $C$, then not only should we be able to conclude that $A$ typically implies $B$ and $C$, we should be able to conclude it *by the same token*. We will spend much of this section formalising this idea and considering its relationship to weak justification.

First, a brief word on our choice of notation in this section. We will tend to use generic defeasible entailment $\approx$ rather than notation denoting a specific formalism thereof such as $\approx_{RC}$. This is partly for typographical compactness and partly because using the more specific syntax distracts from the fact that these results can likely be applied to other notions of KLM-style defeasible entailment (as a matter of fact, we do exactly that for Relevant Closure in Section B.1). In a strict formal sense, however, this presentation is only applicable to Rational Closure.

We present results for each postulate but for the sake of illustration let us begin with a case study of *And*:

*Example 4.2.* Consider the knowledge base $\mathcal{K}$ of the following statements:

(1) Penguins are typically birds ($p \mathrel{|\!\sim} b$).
(2) Penguins typically do not fly ($p \mathrel{|\!\sim} \neg f$).
(3) Birds typically fly ($b \mathrel{|\!\sim} f$).

The *And* axiom tells us for example that

$$\mathcal{K} \approx p \mathrel{|\!\sim} b, \mathcal{K} \approx p \mathrel{|\!\sim} \neg f \implies \mathcal{K} \approx p \mathrel{|\!\sim} b \wedge \neg f.$$

We now consider how the weak justifications of $\mathcal{K} \approx p \mathrel{|\!\sim} b$ and $\mathcal{K} \approx p \mathrel{|\!\sim} \neg f$ relate to the weak justifications of $\mathcal{K} \approx p \mathrel{|\!\sim} b \wedge \neg f$. The justifications here are straightforward, namely:

- $\mathcal{J}_1 = \{p \mathrel{|\!\sim} b\}$ for $\mathcal{K} \approx p \mathrel{|\!\sim} b$,
- $\mathcal{J}_2 = \{p \mathrel{|\!\sim} \neg f\}$ for $\mathcal{K} \approx p \mathrel{|\!\sim} \neg f$, and
- $\mathcal{J}_3 = \{p \mathrel{|\!\sim} b, p \mathrel{|\!\sim} \neg f\}$ for $\mathcal{K} \approx p \mathrel{|\!\sim} b \wedge \neg f$.

Although we have $\mathcal{J}_3 = \mathcal{J}_1 \cup \mathcal{J}_2$ in the example above, counterexamples show that this is not necessarily the case, i.e. if

$$\mathcal{J}_1 \in \mathcal{J}_W(\mathcal{K}, A \mathrel{|\!\sim} B), \mathcal{J}_2 \in \mathcal{J}_W(\mathcal{K}, A \mathrel{|\!\sim} C)$$

then we do not necessarily have $\mathcal{J}_1 \cup \mathcal{J}_2 \in \mathcal{J}_W(\mathcal{K}, A \mathrel{|\!\sim} B \wedge C)$. However, while the resulting unions are not necessarily weak justifications for $\mathcal{K} \approx A \wedge B \mathrel{|\!\sim} C$ (they are not necessarily minimal), we find that they do have interesting properties with respect to the entailment. We will refer to such knowledge bases as *deciding* for an entailment:

*Definition 4.3.* A knowledge base $\mathcal{D} \subseteq \mathcal{K}$ is deciding for an entailment $\mathcal{K} \approx A \mathrel{|\!\sim} B$ if

$$\mathcal{D} \subseteq \mathcal{E}_A^{\mathcal{K}} \text{ and } \mathcal{D} \models A \rightarrow B.$$

[5] This definition is stronger than Lehmann and Magidor's description because we have $\mathcal{K} \approx A \leftrightarrow B$ instead of $A \equiv B$ in *LLE* and $\mathcal{K} \approx A \rightarrow B$ instead of $A \models B$ in *RW*. This provides a more interesting perspective for our purposes.

For a given entailment, any deciding knowledge base is always a superset of a weak justification and all weak justifications are deciding (refer to Proposition 3.3). We also have the following results for deciding knowledge bases:

Proposition 4.4. *If $\mathcal{D}$ is a deciding knowledge base for an entailment $\mathcal{K} \approx A \mid\sim B$ and $\mathrm{br}^{\mathcal{K}}(A) \neq \infty$, then $\mathcal{D} \approx A \mid\sim B$.*

Proposition 4.5. *If $\mathcal{D}$ is a deciding knowledge base for an entailment $\mathcal{K} \approx A \mid\sim B$ and we have $\mathcal{D} \approx A \mid\sim B$, then*

$$\mathcal{J}_W(\mathcal{D}, A \mid\sim B) \subseteq \mathcal{J}_W(\mathcal{K}, A \mid\sim B).$$

Proofs for these propositions, as well as all results in this section, are given in Appendix A. We can now state our result for *And* formally as well as similar results for three other axioms of rational entailment:

Theorem 4.6. *For any knowledge bases $\mathcal{K}, \mathcal{J}_1, \mathcal{J}_2$,*

*(LLE)* if $\mathcal{J}_1 \in \mathcal{J}_W(\mathcal{K}, A \leftrightarrow B)$ and $\mathcal{J}_2 \in \mathcal{J}_W(\mathcal{K}, A \mid\sim C)$, then $\mathcal{J}_1 \cup \mathcal{J}_2$ is deciding for $\mathcal{K} \approx B \mid\sim C$;

*(RW)* if $\mathcal{J}_1 \in \mathcal{J}_W(\mathcal{K}, A \rightarrow B)$ and $\mathcal{J}_2 \in \mathcal{J}_W(\mathcal{K}, C \mid\sim A)$, then $\mathcal{J}_1 \cup \mathcal{J}_2$ is deciding for $\mathcal{K} \approx C \mid\sim B$;

*(And)* if $\mathcal{J}_1 \in \mathcal{J}_W(\mathcal{K}, A \mid\sim B)$ and $\mathcal{J}_2 \in \mathcal{J}_W(\mathcal{K}, A \mid\sim C)$, then $\mathcal{J}_1 \cup \mathcal{J}_2$ is deciding for $\mathcal{K} \approx A \mid\sim B \wedge C$;

*(Or)* if $\mathcal{J}_1 \in \mathcal{J}_W(\mathcal{K}, A \mid\sim C)$ and $\mathcal{J}_2 \in \mathcal{J}_W(\mathcal{K}, B \mid\sim C)$, then $\mathcal{J}_1 \cup \mathcal{J}_2$ is deciding for $\mathcal{K} \approx A \vee B \mid\sim C$.

The propositions and definition above convey that the resulting unions of the form $\mathcal{J}_1 \cup \mathcal{J}_2$ do contain sufficient information that the resulting statement can be drawn as a entailment of $\mathcal{J}_1 \cup \mathcal{J}_2$ itself. This is true both defeasibly (Proposition 4.4) and for the materialisation (Definition 4.3). We can also easily identify the weak justifications with respect to $\mathcal{K}$ contained within $\mathcal{J}_1 \cup \mathcal{J}_2$ via Proposition 4.5. The caveat here is that this does not necessarily apply if $\mathrm{br}^{\mathcal{K}}(A) = \infty$ where $A$ is the antecedent of the resulting statement; this is an interesting characteristic of weak justification which we discuss later.

Having considered the first four postulates for rationality, we now turn our attention to the remaining *Ref*, *CM* and *RM* which are unlike the other postulates and require a different approach. First, notice that *Ref* is the only axiom that has no preconditions and allows for deductions to by made even for the empty knowledge base. We find a related property characteristic of weak justification which states that entailment queries of the form $A \mid\sim A$ are justified solely by the empty set:

Theorem 4.7. *For any knowledge base $\mathcal{K}$ and propositional formula $A$,*

*(Ref)* $\mathcal{J}_W(\mathcal{K}, A \mid\sim A) = \{\emptyset\}$.

The intuitive interpretation of this property is that entailments of the form $\mathcal{K} \approx A \mid\sim A$ are self-evident and that justifications of such entailments do not contain extraneous information, i.e. they are minimal.

Lastly we consider *CM* and *RM.* Note that because *CM* is a strictly weaker condition than *RM*, we do not really need to consider *CM*; however, since there may be analytical value in considering *CM* independently of *RM,* we state results for both. The *RM* postulate is unique in that it specifies a particular entailment is a consequence of a prior entailment *in the absence* of more specific information to the contrary. Our adaptation here states that the justifications of the prior entailment are deciding for the consequent entailment in the absence of such information:

Theorem 4.8. *For any knowledge base $\mathcal{K}$,*

*(CM)* if $\mathcal{K} \approx A \mid\sim C$ and $\mathcal{K} \approx A \mid\sim B$, then every $\mathcal{J} \in \mathcal{J}_W(\mathcal{K}, A \mid\sim C)$ is deciding for $\mathcal{K} \approx A \wedge B \mid\sim C$;

*(RM)* if $\mathcal{K} \approx A \mid\sim C$ and $\mathcal{K} \not\approx A \mid\sim \neg B$, then every $\mathcal{J} \in \mathcal{J}_W(\mathcal{K}, A \mid\sim C)$ is deciding for $\mathcal{K} \approx A \wedge B \mid\sim C$.

As an aside, we mentioned earlier that it was curious that Propositions 4.4 and 4.5 do not hold when $\mathrm{br}^{\mathcal{K}}(A) = \infty$ where $A$ is the antecedent of the query $A \mid\sim B$. This is due to pathological cases such as the following:

*Example 4.9.* Consider the entailment $\mathcal{K} \approx p \mid\sim x$ for

$$\mathcal{K} = \{p \mid\sim b \wedge \neg f, p \mid\sim x, p \mid\sim \bot, b \mid\sim f, b \mid\sim \bot\}.$$

It should be easy to see that every statement is in $\mathcal{E}_\infty^{\mathcal{K}}$, i.e. $\mathcal{K} = \mathcal{E}_\infty^{\mathcal{K}}$. A deciding knowledge base for $\mathcal{K} \approx p \mid\sim x$ is

$$\mathcal{D} = \{p \mid\sim b \wedge \neg f, b \mid\sim f\}$$

since $\overline{\mathcal{D}} \models \neg p \implies \overline{\mathcal{D}} \models p \rightarrow x$. However, we do not have that $\mathcal{D} \approx p \mid\sim x$. Therefore the result in Proposition 4.4 does not necessarily apply if the base rank of the antecedent is $\infty$.

There is some suggestion here that perhaps the way we have defined weak justification (and deciding knowledge bases) for cases where $\mathrm{br}^{\mathcal{K}}(A) = \infty$ is not ideal. In a defeasible context, we would not ordinarily think of the pair of statements $p \mid\sim b \wedge \neg f, b \mid\sim f$ as sufficient to conclude that $\neg p$ (then having $\neg p \models p \rightarrow x$) given antecedent $p$. We discuss this matter in more detail in Appendix B (Section B.2).

We end this section with some comments about the the usefulness of these results. The main value of these properties is that they offer intuitive evidence that weak justification is overall a sensible way to justify KLM-style defeasible entailment. Recall the intuition we expressed at the beginning of this section: that if the implications of $A$ and $B$ are justified in a particular way, the implication of $A \wedge B$ ought to be justified by the same token. Although they were not identified *a priori,* the theorems in this section, together with Propositions 4.4 and 4.5, appear to formalise this idea and offer evidence that such intuitions are true of weak justification. We find results not only for *And* but for each of the axioms of rational entailment, each with a corresponding intuitive interpretation. These results also identify that weak justification is a weakening of classical justification in a way reminiscent of how defeasible entailment is a weakening of classical entailment;[6] this is desirable because intuitively speaking we expect defeasible reasoning to inherit many properties of classical reasoning apart from of course monotonicity. Finally, we suggest that these theorems may offer a perspective from which to analyse not just weak justification but potentially other notions of defeasible justification as identifying which analogous theorems hold for other notions may reveal intuitive differences between justification formalisms that would otherwise be difficult to grasp.

[6]Although we have not provided proofs, it is trivial to show that classical justification satisfies similar properties for classical logic.

## 5 WEAK EXPLANATION FOR RELEVANT CLOSURE

We noted earlier that weak justification has only been explored for the case of Rational Closure entailment. In this section, we adapt this result to the case of Relevant Closure. As discussed in Section 3.1, the essential principle behind weak justification is that we first eliminate more general statements and then make use of classical justification on the remaining statements by materialising. The difference for Relevant Closure is that unlike Rational Closure we do not eliminate more general statements that are not considered relevant to the query. An analogue of weak justification (Proposition 3.3) for the case of Relevant Closure entailment is then the following:

*Definition 5.1.* A knowledge base $\mathcal{J}$ is a weak justification for a Relevant Closure entailment $\mathcal{K} \approx_{\text{BRelC}} A \mathrel{|\!\sim} B$ or $\mathcal{K} \approx_{\text{MRelC}} A \mathrel{|\!\sim} B$ if $\overline{\mathcal{J}}$ is a classical justification for

$$\overline{\mathcal{E}_A^{\mathcal{K}} \cup R^-} \models A \rightarrow B$$

where $(R, R^-)$ is the relevance partition of $\mathcal{K}$ for $A \mathrel{|\!\sim} B$ given the form of relevance at hand.

*Example 5.2.* Consider the knowledge base $\mathcal{K}$ from Example 2.16:

$$\mathcal{K} = \{b \mathrel{|\!\sim} f, b \mathrel{|\!\sim} w, p \rightarrow b, p \mathrel{|\!\sim} \neg f, r \rightarrow p\}\,.$$

Suppose we wish to find the weak justifications for the entailment $\mathcal{K} \approx_{\text{MRelC}} r \mathrel{|\!\sim} w$. As we had in Example 2.27,

$$\mathcal{E}_r^{\mathcal{K}} \cup R^- = \{p \mathrel{|\!\sim} \neg f, p \rightarrow b, r \rightarrow p, b \mathrel{|\!\sim} w\}\,.$$

Then applying the definition above gives us that a unique weak justification for $\mathcal{K} \approx_{\text{MRelC}} r \mathrel{|\!\sim} w$ is

$$\{r \rightarrow p, p \rightarrow b, b \mathrel{|\!\sim} w\}\,.$$

In other words, we ensure that the statements in the knowledge base considered not relevant to the query remain under our consideration when materialising just as they are when evaluating Relevant Closure entailment queries (Definition 2.26). The results here seem to be generally intuitive and correspond to the definition of defeasible entailment as we had in the case of Rational Closure. We can compute these justifications by adapting Algorithms 2 and 4. This result is given in Algorithm 5 and proved in Appendix A (Proposition A.4):

**ALGORITHM 5:** `WeakJustifyRelC`

**Data:** A knowledge base $\mathcal{K}$, query $A \mathrel{|\!\sim} B$ and a relevance partition $(R, R^-)$
**Result:** The weak justifications for $\mathcal{K} \approx_{\text{BRelC}} A \mathrel{|\!\sim} B$ or $\mathcal{K} \approx_{\text{MRelC}} A \mathrel{|\!\sim} B$ (depending on the relevance partition)

$(\mathcal{K}_0, \cdots, \mathcal{K}_n, n) := \mathsf{Rank}(\mathcal{K})$
$i := 0$
$\mathcal{K}' := \mathcal{K}$
**while** $i < n$ *and* $\overline{\mathcal{K}'} \models \neg A$ **do**
  $\mathcal{K}' := \mathcal{K}' \setminus (\mathcal{K}_i \cap R)$
  $i := i + 1$
**end**
**return** $\left\{\underline{\mathcal{J}} \mid \mathcal{J} \in \mathcal{J}\left(\overline{\mathcal{K}'}, A \rightarrow B\right)\right\}$

It is useful to analyse this result in context of the previous section. We noted in our introduction of Relevant Closure that it is not rational, but interestingly, Relevant Closure does satisfy *some* of the postulates for rationality, namely *LLE, RW, And* and *Ref* [4]. This means that even though Relevant Closure is not rational, it is possible to analyse Relevant Closure with respect to the properties discussed in Section 4. These results are presented in Appendix B (Section B.1).

## 6 RELATED WORK

Kraus, Lehmann and Magidor [14] describe the KLM approach to defeasible reasoning. Lehmann and Magidor [17] describe Rational Closure, the postulates for rational defeasible entailment and an algorithm for deciding Rational Closure entailment. Kaliski [13] compiles the literature on the topic and offers a contemporary summary of the work on KLM-style defeasible reasoning.

Horridge [12] presents a detailed look at classical justification and its computation and proposes algorithms for enumerating classical justifications. Casini et al. [4] introduce Relevant Closure, a notion of defeasibility for KLM based on classical justification. Chama [9] proposes a notion of defeasible justification by describing a justification algorithm for Rational Closure. Brewka and Ulbricht [3] on the other hand propose *strong explanation* using a declarative definition which may be applicable to other frameworks of defeasible reasoning.

## 7 CONCLUSIONS

Our analysis suggests that weak justification has a much closer correspondence to KLM-style defeasible entailment than does strong justification. We were able to characterise weak justification in terms of declarative properties with intuitive interpretations by adapting the postulates for rationality given by Lehmann and Magidor [17]. This result suggests that weak justification, except perhaps for a boundary case which we attempt to resolve, is a sound notion of justification for KLM-style defeasible entailment. This result also shows similarities between weak justification for the defeasible case and classical justification for the classical case, and may offer a perspective from which to evaluate other notions of defeasible justification. Our final result is an adaptation of weak justification from Rational Closure to Relevant Closure. Here, we show that weak justification has a direct analogue for Relevant Closure, suggest an algorithm that enumerates these weak justifications for and show that the properties identified earlier similarly apply here.

## 8 FUTURE WORK

Since our exploration was limited to the case of Rational and Relevant Closure, further work might identify a definition of weak justification applicable to all rational entailment relations using a more abstract view of reasoning algorithms such as that given by Casini, Meyer and Varzinczak [6]. One could potentially then prove that theorems similar to those in Section 4 hold for all rational entailment relations. Another possibility is to consider how these ideas apply to KLM description logic [5, 8, 18]. Lastly, one could evaluate notions of defeasible explanation for KLM beyond the ideas discussed in this paper such as perhaps the revised definition of strong justification proposed in Section 3.2.

## ACKNOWLEDGMENTS

My sincere thanks to our supervisor Tommie Meyer for his patience and insight and to my colleague Emily Morris for many interesting discussions about KLM and otherwise.


## REFERENCES


[1] S. P. Bail. 2013. *The justificatory structure of OWL ontologies.* Ph.D. Dissertation. University of Manchester.

[2] Or Biran and Courtenay Cotton. 2017. Explanation and justification in machine learning: A survey. In *IJCAI-17 workshop on explainable AI (XAI)*, Vol. 8. 8–13.

[3] Gerhard Brewka and Markus Ulbricht. 2019. *Strong Explanations for Nonmonotonic Reasoning.* Springer International Publishing, Cham, 135–146. https://doi.org/10.1007/978-3-030-22102-7_6

[4] Giovanni Casini, Thomas Meyer, Kodylan Moodley, and Riku Nortjé. 2014. Relevant Closure: A New Form of Defeasible Reasoning for Description Logics. In *Logics in Artificial Intelligence*, Eduardo Fermé and João Leite (Eds.). Springer International Publishing, Cham, 92–106.

[5] Giovanni Casini, Thomas Meyer, Kodylan Moodley, and Ivan Varzinczak. 2013. Towards practical defeasible reasoning for description logics. (2013).

[6] Giovanni Casini, Thomas Meyer, and Ivan Varzinczak. 2019. Taking Defeasible Entailment Beyond Rational Closure. In *Logics in Artificial Intelligence*, Francesco Calimeri, Nicola Leone, and Marco Manna (Eds.). Springer International Publishing, Cham, 182–197.

[7] Giovanni Casini, Thomas Meyer, and Ivan Varzinczak. 2019. Taking Defeasible Entailment Beyond Rational Closure. In *Logics in Artificial Intelligence*, Francesco Calimeri, Nicola Leone, and Marco Manna (Eds.). Springer International Publishing, Cham, 182–197.

[8] Giovanni Casini and Umberto Straccia. 2013. Defeasible inheritance-based description logics. *Journal of Artificial Intelligence Research* 48 (2013), 415–473.

[9] V. Chama. 2020. *Explanation for defeasible entailment.* Master's thesis. University of Cape Town, Cape Town, South Africa. http://hdl.handle.net/11427/32206

[10] Michael Freund. 1998. Preferential Reasoning in the Perspective of Poole Default Logic. *Artificial Intelligence* 98, 1–2 (Jan. 1998), 209–235. https://doi.org/10.1016/S0004-3702(97)00053-2

[11] L. Giordano, V. Gliozzi, N. Olivetti, and G.L. Pozzato. 2015. Semantic characterization of rational closure: From propositional logic to description logics. *Artificial Intelligence* 226 (2015), 1–33. https://doi.org/10.1016/j.artint.2015.05.001

[12] Matthew Horridge. 2011. *Justification Based Explanation in Ontologies.* Ph.D. Dissertation. University of Manchester.

[13] A. Kaliski. 2020. *An overview of KLM-style defeasible entailment.* Master's thesis. University of Cape Town, Cape Town, South Africa. http://hdl.handle.net/11427/32743

[14] Sarit Kraus, Daniel Lehmann, and Menachem Magidor. 1990. Nonmonotonic reasoning, preferential models and cumulative logics. *Artificial Intelligence* 44, 1 (1990), 167–207. https://doi.org/10.1016/0004-3702(90)90101-5

[15] Markus Krötzsch and Daria Stepanova. 2019. *Reasoning Web. Explainable Artificial Intelligence: 15th International Summer School 2019, Bolzano, Italy, September 20–24, 2019, Tutorial Lectures.* Vol. 11810. Springer Nature.

[16] Daniel Lehmann. 1995. Another perspective on default reasoning. *Annals of mathematics and artificial intelligence* 15, 1 (1995), 61–82.

[17] Daniel Lehmann and Menachem Magidor. 1992. What does a conditional knowledge base entail? *Artificial Intelligence* 55, 1 (1992), 1–60. https://doi.org/10.1016/0004-3702(92)90041-U

[18] Kody Moodley. 2016. *Practical reasoning for defeasable description logics.* Ph.D. Dissertation. University of KwaZulu-Natal.

# A PROOFS

## A.1 Properties for Defeasible Explanation (Section 4)

As discussed in Section 4, we only consider the case of Rational Closure entailment, denoted here as $\approx$. We first introduce some lemmas which will help us with the proofs in this section:

Lemma A.1. *If $\mathcal{K} \approx A \to B$ for a knowledge base $\mathcal{K}$ and propositional formulas $A, B$,*

$$\mathrm{br}^{\mathcal{K}}(B) \leq \mathrm{br}^{\mathcal{K}}(A).$$

Proof. Let $\mathcal{K}^* = \overline{\mathcal{E}_A^{\mathcal{K}}}$. If $\mathrm{br}^{\mathcal{K}}(A) = \infty$ then trivially the condition above is satisfied. Otherwise, suppose by contradiction that $\mathrm{br}^{\mathcal{K}}(B) > \mathrm{br}^{\mathcal{K}}(A)$. Clearly $\mathcal{K}^* \not\models \neg A$ and $\mathcal{K}^* \models \neg B$. Consider any model $\mathcal{I}$ of $\mathcal{K}^*$ such that $\mathcal{I}(A) = \mathrm{T}$. Since $\mathcal{K}^* \models A \to B$ by assumption, every $\mathcal{I}$ has $\mathcal{I}(B) = \mathrm{T}$. However, since $\mathcal{K}^* \models \neg B$ there can be no such $\mathcal{I}$. This means that $\mathcal{K}^* \models \neg A$ and we have a contradiction. □

Corollary A.2. *If $A \models B$ then for any knowledge base $\mathcal{K}$ and propositional formulas $A, B$,*

$$\mathrm{br}^{\mathcal{K}}(B) \leq \mathrm{br}^{\mathcal{K}}(A).$$

Proof. This result follows from Lemma A.1 above since $A \models B$ implies that $A \to B$ is a tautology. □

This result formalises the intuition that for example $A \wedge B$ cannot be any less specific than either $A$ or $B$: applying Lemma A.2 to $A \wedge B \models A$ and $A \wedge B \models B$ gives that for any knowledge base $\mathcal{K}$,

$$\max\left(\mathrm{br}^{\mathcal{K}}(A), \mathrm{br}^{\mathcal{K}}(B)\right) \leq \mathrm{br}^{\mathcal{K}}(A \wedge B).$$

We have a similar but inverse result for disjunctions where $A \vee B$ cannot be any more specific than either $A$ or $B$.

We now point out that weak justifications of classical statements are always subsets of $\mathcal{E}_\infty^{\mathcal{K}}$ for the knowledge base $\mathcal{K}$ in question:

Lemma A.3. *For a propositional formula $A$ and knowledge base $\mathcal{K}$, every $\mathcal{J} \in \mathcal{J}_W(\mathcal{K} \approx A)$ has $\mathcal{J} \subseteq \mathcal{E}_\infty^{\mathcal{K}}$.*

Proof. In this context $A$ is a shorthand for $\neg A \mathrel{|\!\sim} \bot$. The base rank $\mathrm{br}^{\mathcal{K}}(\neg A)$ of the antecedent is the smallest $i$ with $0 \leq i \leq n$ such that $\neg A$ is not exceptional for $\mathcal{E}_i^{\mathcal{K}}$ or otherwise $\infty$ if there is no such $i$. There can be no such $i$ because if there were we would have

$$\mathrm{br}^{\mathcal{K}}(\neg A) = i \text{ and } \overline{\mathcal{E}_i^{\mathcal{K}}} \not\models \neg A \to \bot,$$

which would mean that $\mathcal{K} \not\approx A$ (i.e. $\mathcal{K} \not\approx \neg A \mathrel{|\!\sim} \bot$). Therefore $\mathrm{br}^{\mathcal{K}}(\neg A) = \infty$. It follows easily that any weak justification $\mathcal{J}$ of $\mathcal{K} \approx A$ has $\mathcal{J} \subseteq \mathcal{E}_\infty^{\mathcal{K}}$ (refer to Proposition 3.3). □

We can now prove our results in Section 4:

Proposition 4.4. *If $\mathcal{D}$ is a deciding knowledge base for an entailment $\mathcal{K} \approx A \mathrel{|\!\sim} B$ and $\mathrm{br}^{\mathcal{K}}(A) \neq \infty$, then $\mathcal{D} \approx A \mathrel{|\!\sim} B$.*

Proof. Since $\mathrm{br}^{\mathcal{K}}(A) \neq \infty$, we have $\overline{\mathcal{E}_A^{\mathcal{K}}} \not\models \neg A$. Then by classical monotonicity (or rather its contrapositive) we have $\overline{\mathcal{D}} \not\models \neg A$. Hence $\mathcal{E}_A^{\mathcal{D}} = \mathcal{D}$ and because $\mathcal{D}$ is deciding,

$$\overline{\mathcal{E}_A^{\mathcal{D}}} \models A \to B$$

which implies that $\mathcal{D} \approx A \mathrel{|\!\sim} B$. □

Proposition 4.5. *If $\mathcal{D}$ is a deciding knowledge base for an entailment $\mathcal{K} \approx A \mathrel{|\!\sim} B$ and we have $\mathcal{D} \approx A \mathrel{|\!\sim} B$, then $\mathcal{J}_W(\mathcal{D}, A \mathrel{|\!\sim} B) \subseteq \mathcal{J}_W(\mathcal{K}, A \mathrel{|\!\sim} B)$.*

Proof. Since $\mathcal{D}$ is deciding, we have $\mathcal{D} \subseteq \mathcal{E}_A^{\mathcal{K}}$ and $\overline{\mathcal{D}} \models A \to B$. All $\mathcal{J} \in \mathcal{J}_W(\mathcal{D}, A \mathrel{|\!\sim} B)$ have $\mathcal{J} \subseteq \mathcal{D}$ and $\overline{\mathcal{J}} \models A \to B$. Each $\mathcal{J}$ is also minimal; for every $\mathcal{J}$ there is no $\mathcal{J}' \subset \mathcal{J}$ with $\mathcal{J}' \models A \to B$. Therefore every $\mathcal{J}$ has $\mathcal{J} \in \mathcal{J}_W(\mathcal{K}, A \mathrel{|\!\sim} B)$. □

Theorem 4.6. *For any knowledge bases $\mathcal{K}, \mathcal{J}_1, \mathcal{J}_2$,*

- *(LLE) if $\mathcal{J}_1 \in \mathcal{J}_W(\mathcal{K}, A \leftrightarrow B)$ and $\mathcal{J}_2 \in \mathcal{J}_W(\mathcal{K}, A \mathrel{|\!\sim} C)$, then $\mathcal{J}_1 \cup \mathcal{J}_2$ is deciding for $\mathcal{K} \approx B \mathrel{|\!\sim} C$;*
- *(RW) if $\mathcal{J}_1 \in \mathcal{J}_W(\mathcal{K}, A \to B)$ and $\mathcal{J}_2 \in \mathcal{J}_W(\mathcal{K}, C \mathrel{|\!\sim} A)$, then $\mathcal{J}_1 \cup \mathcal{J}_2$ is deciding for $\mathcal{K} \approx C \mathrel{|\!\sim} B$;*
- *(And) if $\mathcal{J}_1 \in \mathcal{J}_W(\mathcal{K}, A \mathrel{|\!\sim} B)$ and $\mathcal{J}_2 \in \mathcal{J}_W(\mathcal{K}, A \mathrel{|\!\sim} C)$, then $\mathcal{J}_1 \cup \mathcal{J}_2$ is deciding for $\mathcal{K} \approx A \mathrel{|\!\sim} B \wedge C$;*
- *(Or) if $\mathcal{J}_1 \in \mathcal{J}_W(\mathcal{K}, A \mathrel{|\!\sim} C)$ and $\mathcal{J}_2 \in \mathcal{J}_W(\mathcal{K}, B \mathrel{|\!\sim} C)$, then $\mathcal{J}_1 \cup \mathcal{J}_2$ is deciding for $\mathcal{K} \approx A \vee B \mathrel{|\!\sim} C$.*

Proof. We consider each statement in sequence:

- (LLE) Applying Lemma A.3 gives that $\mathcal{J}_1 \subseteq \mathcal{E}_\infty^{\mathcal{K}}$. Also notice that it follows from $\mathcal{K} \approx A \leftrightarrow B$ that $\mathcal{K} \approx A \to B$ and $\mathcal{K} \approx B \to A$.[7] Applying Lemma A.1 then gives $\mathrm{br}^{\mathcal{K}}(A) = \mathrm{br}^{\mathcal{K}}(B)$. Since $\mathcal{J}_1$ and $\mathcal{J}_2$ are weak justifications and $\mathcal{J}_1 \subseteq \mathcal{E}_\infty^{\mathcal{K}}$, we have $\overline{\mathcal{J}_1} \models A \leftrightarrow B$, $\overline{\mathcal{J}_2} \models A \to C$ and $\mathcal{J}_1 \cup \mathcal{J}_2 \subseteq \mathcal{E}_A^{\mathcal{K}}$. Then $\overline{\mathcal{J}_1 \cup \mathcal{J}_2} \models B \to C$ and $\mathcal{J}_1 \cup \mathcal{J}_2 \subseteq \mathcal{E}_B^{\mathcal{K}}$; hence $\mathcal{J}_1 \cup \mathcal{J}_2$ is deciding for $\mathcal{K} \approx B \mathrel{|\!\sim} C$.
- (RW) Applying Lemma A.3 gives that $\mathcal{J}_1 \subseteq \mathcal{E}_\infty^{\mathcal{K}}$. We then have $\overline{\mathcal{J}_1} \models A \to B$, $\overline{\mathcal{J}_2} \models C \to A$ and $\mathcal{J}_1 \cup \mathcal{J}_2 \subseteq \mathcal{E}_C^{\mathcal{K}}$. Therefore $\overline{\mathcal{J}_1 \cup \mathcal{J}_2} \models C \to B$ and $\mathcal{J}_1 \cup \mathcal{J}_2$ is deciding for $\mathcal{K} \approx C \mathrel{|\!\sim} B$.
- (And) We have $\overline{\mathcal{J}_1} \models A \to C$, $\overline{\mathcal{J}_2} \models A \to B$ and $\mathcal{J}_1 \cup \mathcal{J}_2 \subseteq \mathcal{E}_A^{\mathcal{K}}$. We then have $\overline{\mathcal{J}_1 \cup \mathcal{J}_2} \models A \to C \wedge B$ and therefore $\mathcal{J}_1 \cup \mathcal{J}_2$ is deciding for $\mathcal{K} \approx A \mathrel{|\!\sim} B \wedge C$.
- (Or) Applying Corollary A.2 to $A \models A \vee B$ and $B \models A \vee B$ gives that $\mathrm{br}^{\mathcal{K}}(A \vee B) \leq \mathrm{br}^{\mathcal{K}}(A)$ and $\mathrm{br}^{\mathcal{K}}(A \vee B) \leq \mathrm{br}^{\mathcal{K}}(B)$. We have $\overline{\mathcal{J}_1} \models A \to C$, $\overline{\mathcal{J}_2} \models B \to C$ and $\mathcal{J}_1 \cup \mathcal{J}_2 \subseteq \mathcal{E}_A^{\mathcal{K}} \cup \mathcal{E}_B^{\mathcal{K}}$. It follows that $\overline{\mathcal{J}_1 \cup \mathcal{J}_2} \models A \vee B \to C$ and $\mathcal{J}_1 \cup \mathcal{J}_2 \subseteq \mathcal{E}_{A \vee B}^{\mathcal{K}}$; therefore $\mathcal{J}_1 \cup \mathcal{J}_2$ is deciding for $\mathcal{K} \approx A \vee B \mathrel{|\!\sim} C$.

□

Theorem 4.7. *For any knowledge base $\mathcal{K}$ and propositional formula $A$,*

*(Ref)* $\mathcal{J}_W(\mathcal{K}, A \mathrel{|\!\sim} A) = \{\emptyset\}$.

[7] The full deduction here is $\mathcal{K} \approx A \leftrightarrow B \implies \overline{\mathcal{E}_\infty^{\mathcal{K}}} \models \neg(A \leftrightarrow B) \to \bot \implies \overline{\mathcal{E}_\infty^{\mathcal{K}}} \models \neg(A \to B) \to \bot \implies \mathcal{K} \approx A \to B$ and likewise for $\mathcal{K} \approx B \to A$.

Proof. Clearly $\emptyset$ is a weak justification for $\mathcal{K} \approx A \vdash A$ since $\emptyset \subseteq \mathcal{E}_A^{\mathcal{K}}$ and $\emptyset \models A \rightarrow A$. There are no other weak justifications because any $\mathcal{K}' \neq \emptyset$ is not minimal. ◻

Theorem 4.8. *For any knowledge base $\mathcal{K}$,*

(CM) *if $\mathcal{K} \approx A \vdash C$ and $\mathcal{K} \approx A \vdash B$, then every $\mathcal{J} \in \mathcal{J}_W(\mathcal{K}, A \vdash C)$ is deciding for $\mathcal{K} \approx A \wedge B \vdash C$;*

(RM) *if $\mathcal{K} \approx A \vdash C$ and $\mathcal{K} \not\approx A \vdash \neg B$, then every $\mathcal{J} \in \mathcal{J}_W(\mathcal{K}, A \vdash C)$ is deciding for $\mathcal{K} \approx A \wedge B \vdash C$.*

Proof. Consider any $\mathcal{J} \in \mathcal{J}_W(\mathcal{K}, A \vdash C)$. We have $\overline{\mathcal{J}} \models A \rightarrow C$ and therefore $\overline{\mathcal{J}} \models A \wedge B \rightarrow C$. Since $\mathcal{K} \not\approx A \vdash \neg B$, $B$ is not exceptional for $\mathcal{E}_A^{\mathcal{K}}$, i.e. $\text{br}^{\mathcal{K}}(B) \leq \text{br}^{\mathcal{K}}(A)$. Therefore $\text{br}^{\mathcal{K}}(A \wedge B) \geq \text{br}^{\mathcal{K}}(A)$ by Corollary A.2. We now prove $\text{br}^{\mathcal{K}}(A \wedge B) = \text{br}^{\mathcal{K}}(A)$. Suppose by contradiction that $\text{br}^{\mathcal{K}}(A \wedge B) > \text{br}^{\mathcal{K}}(A)$. For compactness, let

$$\mathcal{K}^* = \overline{\mathcal{E}_A^{\mathcal{K}}}.$$

Then $\mathcal{K}^* \models \neg A \vee \neg B$ but also $\mathcal{K}^* \models \neg A$ since $\text{br}^{\mathcal{K}}(A) \neq \infty$. This means that $\mathcal{K}^* \models \neg B$ but this is a contradiction since earlier we had that $\mathcal{K}^* \not\models \neg B$. Therefore $\text{br}^{\mathcal{K}}(A \wedge B) = \text{br}^{\mathcal{K}}(A)$. This implies that $\mathcal{J} \subseteq \mathcal{E}_{A \wedge B}^{\mathcal{K}}$, and since $\overline{\mathcal{J}} \models A \wedge B \rightarrow C$, we have that $\overline{\mathcal{J}}$ is deciding for $\mathcal{K} \approx A \wedge B \vdash C$. ◻

## A.2 Weak Explanation for Relevant Closure (Section 5)

Proposition A.4. *A knowledge base $\mathcal{J}$ is a weak justification for a Relevant Closure entailment $\mathcal{K} \approx_{\text{BRelC}} A \vdash B$ or $\mathcal{K} \approx_{\text{MRelC}} A \vdash B$ if it is returned by Algorithm 5 given $\mathcal{K}, A \vdash B$ and the corresponding relevance partition $(R, R^-)$.*

Proof. As we reach the return statement in Algorithm 5 we have $\mathcal{K}' = \mathcal{E}_A^{\mathcal{K}} \cup R$ since up to this statement the algorithm is the same as Algorithm 2. It easily follows that

$$\left\{ \underline{\mathcal{J}} \mid \mathcal{J} \in \mathcal{J}\left(\overline{\mathcal{K}'}, A \rightarrow B\right)\right\}$$

is the set of weak justifications for the form of Relevant Closure at hand since taking the defeasible counterpart is the inverse of materialisation. ◻

# B FURTHER DISCUSSION

## B.1 Properties of Weak Justification for Relevant Closure

We noted earlier that Relevant Closure obeys *LLE*, *RW*, *And* and *Ref*. Using the definition of weak justification for Relevant Closure given in Section 5, we can show that (with a minor adjustment to *LLE*) the properties corresponding to these axioms in Theorems 4.6 and 4.7 hold for Relevant Closure.

For this section, let $\approx$ refer to either Minimal or Basic Relevant Closure entailment. References to weak justification, including the $\mathcal{J}_W(\cdot, \cdot)$ syntax, are to the corresponding Relevant Closure counterpart of weak justification. It is also helpful to introduce some helper syntax for relevance partitions:

*Definition B.1.* For a knowledge base $\mathcal{K}$ and a propositional formula $A$, let $R(\mathcal{K}, A) = R$ and $R^-(\mathcal{K}, A) = R^-$ where $(R, R^-)$ is the relevance partition of $\mathcal{K}$ for $A \vdash B$ (for any $B \in \mathcal{L}$ as relevance is not a function of the consequent) given the form of Relevant Closure at hand.

We also adapt our definition of deciding knowledge bases:

*Definition B.2.* In the context of Relevant Closure, a knowledge base $\mathcal{D}$ is deciding for an entailment $\mathcal{K} \approx A \vdash B$ if

$$\mathcal{D} \subseteq \mathcal{E}_A^{\mathcal{K}} \cup R^-(\mathcal{K}, A) \text{ and } \overline{\mathcal{D}} \models A \rightarrow B.$$

We introduce a lemma similar to Lemma A.3 before presenting our main result:

Lemma B.3. *For a propositional formula $A$ and knowledge base $\mathcal{K}$, every $\mathcal{J} \in \mathcal{J}_W(\mathcal{K} \approx A)$ has $\mathcal{J} \subseteq \mathcal{E}_\infty^{\mathcal{K}}$.*

Proof. Here $A$ is a shorthand for $\neg A \vdash \bot$. We have $\text{br}^{\mathcal{K}}(\neg A) = \infty$ (discussed in the proof for Lemma A.3). Every weak justification $\mathcal{J}$ of $\mathcal{K} \approx A$ then has

$$\overline{\mathcal{J}} \subseteq \mathcal{E}_\infty^{\mathcal{K}} \cup R^-(\mathcal{K}, \neg A) \text{ and } \mathcal{J} \models \neg A \rightarrow \bot.$$

We now show that $\mathcal{J} \subseteq \mathcal{E}_\infty^{\mathcal{K}}$. Suppose by contradiction that there is some $B \vdash C \in \mathcal{J}$ with $B \vdash C \in R^-(\mathcal{K}, \neg A) \setminus \mathcal{E}_\infty^{\mathcal{K}}$. Since $\overline{\mathcal{J}} \models A$ (notice that $\neg A \rightarrow \bot \equiv A$), $B \vdash C$ is part of an $\varepsilon$-justification, namely $\mathcal{J}$, for $(\mathcal{K}, \neg A)$. (For $\approx_{\text{MRelC}}$, if $B \vdash C$ does not have the lowest rank of all statements in $\mathcal{J}$ w.r.t. $\mathcal{K}$, then we can choose a $B \vdash C \in \mathcal{J}$ having this property.) But the set $R^-(\mathcal{K}, \neg A)$ is exactly those statements in $\mathcal{K}$ for which this does not hold; therefore we have a contradiction. ◻

We are then able to obtain results for *RW*, *And* and *Ref*:

Theorem B.4. *For knowledge bases $\mathcal{K}, \mathcal{J}_1, \mathcal{J}_2$,*

(RW) *if $\mathcal{J}_1 \in \mathcal{J}_W(\mathcal{K}, A \rightarrow B)$ and $\mathcal{J}_2 \in \mathcal{J}_W(\mathcal{K}, C \vdash A)$, then $\mathcal{J}_1 \cup \mathcal{J}_2$ is deciding for $\mathcal{K} \approx C \vdash B$;*

(And) *if $\mathcal{J}_1 \in \mathcal{J}_W(\mathcal{K}, A \vdash B)$ and $\mathcal{J}_2 \in \mathcal{J}_W(\mathcal{K}, A \vdash C)$, then $\mathcal{J}_1 \cup \mathcal{J}_2$ is deciding for $\mathcal{K} \approx A \vdash B \wedge C$;*

(Ref) *$\mathcal{J}_W(\mathcal{K}, A \vdash A) = \{\emptyset\}$.*

Proof. We consider each statement in sequence:

(RW) By Lemma B.3, $\mathcal{J}_1 \subseteq \mathcal{E}_\infty^{\mathcal{K}}$. We have $\mathcal{J}_1 \models A \rightarrow B$, $\mathcal{J}_2 \models C \rightarrow A$ and $\mathcal{J}_1 \cup \mathcal{J}_2 \subseteq \mathcal{E}_C^{\mathcal{K}} \cup R^-(\mathcal{K}, C)$ and hence $\mathcal{J}_1 \cup \mathcal{J}_2$ is deciding for $\mathcal{K} \approx C \vdash B$.

(And) We have $\overline{\mathcal{J}_1} \models A \rightarrow C$, $\overline{\mathcal{J}_2} \models A \rightarrow B$ and $\mathcal{J}_1 \cup \mathcal{J}_2 \subseteq \mathcal{E}_A^{\mathcal{K}} \cup R^-(\mathcal{K}, A)$. Therefore $\mathcal{J}_1 \cup \mathcal{J}_2$ is deciding for $\mathcal{K} \approx A \vdash B \wedge C$.

(Ref) We have $\emptyset \models A \rightarrow A$ and $\emptyset \subseteq \mathcal{E}_A^{\mathcal{K}} \cup R(\mathcal{K}, A)$. Therefore $\emptyset$ is a weak justification for $\mathcal{K} \approx A \vdash A$. Any other $\mathcal{K}' \neq \emptyset$ would not be minimal and therefore $\emptyset$ is the sole weak justification for the entailment.

◻

The situation for *LLE* is unlike the three axioms above. In Section 4, we noted that we used a strengthening of the postulates for rationality where *LLE* is given as:

If $\mathcal{K} \approx A \leftrightarrow B$ and $\mathcal{K} \approx A \vdash C$ then $\mathcal{K} \approx B \vdash C$.

This is not true in general for Relevant Closure:

*Example B.5.* Consider $a \vdash c \rightarrow x$ and $b \vdash c \rightarrow x$ for

$$\mathcal{K} = \{a \leftrightarrow b, \top \vdash \neg b \wedge c, c \vdash \neg a \wedge x\}.$$

We give $\mathcal{R}_0^{\mathcal{K}}, \cdots, \mathcal{R}_\infty^{\mathcal{K}}$ in Figure 4. The $\varepsilon$-justifications here are:

| | |
|---|---|
| $\infty$ | $a \leftrightarrow b$ |
| 0 | $\top \mathrel{\vert\!\sim} \neg b \wedge c, c \mathrel{\vert\!\sim} \neg a \wedge x$ |

**Figure 4: Ranking $\mathcal{R}_0^{\mathcal{K}}, \cdots, \mathcal{R}_\infty^{\mathcal{K}}$ for Example B.5**

- $\{\top \mathrel{\vert\!\sim} \neg b \wedge c, c \mathrel{\vert\!\sim} \neg a \wedge x\}, \{\top \mathrel{\vert\!\sim} \neg b \wedge c, a \leftrightarrow b\}$ for $(\mathcal{K}, a)$;
- $\{\top \mathrel{\vert\!\sim} \neg b \wedge c\}$ for $(\mathcal{K}, b)$.

This means that $c \mathrel{\vert\!\sim} \neg a \wedge x$ is not relevant to $b$ but is relevant to $a$ (for either basic or minimal relevance) and hence $\mathcal{K} \approx b \mathrel{\vert\!\sim} c \rightarrow x$ but $\mathcal{K} \not\approx a \mathrel{\vert\!\sim} c \rightarrow x$ despite $\mathcal{K} \approx a \leftrightarrow b$.

This example illustrates an interesting characteristic of Relevant Closure as we would expect the implications of $A$ to match those of $B$ if $\mathcal{K} \approx A \leftrightarrow B$. It seems that in this case the minimality requirement prevents the $\varepsilon$-justifications from containing statements that are not relevant formally but are perhaps intuitively relevant. (Note that we can find simpler examples such as the $\varepsilon$-justifications for $(\mathcal{K}, a)$ and $(\mathcal{K}, b)$ given $\mathcal{K} = \{\top \mathrel{\vert\!\sim} \neg a, (a \leftrightarrow b) \wedge \neg b\}$ though the ramifications for entailments here are not as clear.)

Returning to our discussion of weak justification, while we clearly cannot use the version of *LLE* presented above, we can adapt *LLE* as originally given by Lehmann and Magidor:

If $A \equiv B$ and $\mathcal{K} \approx A \mathrel{\vert\!\sim} C$ then $\mathcal{K} \approx B \mathrel{\vert\!\sim} C$.

This result is given as follows:

Theorem B.6. *For knowledge bases $\mathcal{K}, \mathcal{J}$ and propositional formulas $A, B$,*

*(LLE)* *if $A \equiv B$ then $\mathcal{J}_W(\mathcal{K}, A \mathrel{\vert\!\sim} C) = \mathcal{J}_W(\mathcal{K}, B \mathrel{\vert\!\sim} C)$.*

Proof. Since $A \equiv B$, we can easily show that $\mathrm{br}^{\mathcal{K}}(A) = \mathrm{br}^{\mathcal{K}}(B)$ and $\mathrm{R}^{-}(\mathcal{K}, A) = \mathrm{R}^{-}(\mathcal{K}, B)$. We also note that $\overline{\mathcal{J}} \models A \rightarrow C \iff \overline{\mathcal{J}} \models B \rightarrow C$. Therefore the weak justifications of $\mathcal{K} \approx A \mathrel{\vert\!\sim} C$ are the same as the weak justifications for $\mathcal{K} \approx B \mathrel{\vert\!\sim} C$. □

Lastly, we express Propositions 4.4 and 4.5 for the case of Relevant Closure:

Proposition B.7. *If $\mathcal{D}$ is a deciding knowledge base for an entailment $\mathcal{K} \approx A \mathrel{\vert\!\sim} B$ and $\mathrm{br}^{\mathcal{K}}(A) \neq \infty$, then $\mathcal{D} \approx A \mathrel{\vert\!\sim} B$.*

Proof. Since $\mathrm{br}^{\mathcal{K}}(A) \neq \infty$ we have $\overline{\mathcal{E}_A^{\mathcal{K}}} \not\models \neg A$. Notice also that

$$\overline{\mathcal{E}_A^{\mathcal{K}} \cup R^{-}(\mathcal{K}, A)} \not\models \neg A$$

since the statements in $R(\mathcal{K}, A)$ would be necessary to conclude $\neg A$ (see the proof of Lemma B.3 for this idea in more detail). Since $\mathcal{D}$ is a subset of $\mathcal{E}_A^{\mathcal{K}} \cup R^{-}(\mathcal{K}, A)$, the contrapositive of classical monotonicity implies $\overline{\mathcal{D}} \not\models \neg A$. This implies that $\mathcal{E}_A^{\mathcal{D}} = \mathcal{D}$ and because $\mathcal{D}$ is deciding, $\overline{\mathcal{D}} \models A \rightarrow B$. Therefore $\mathcal{D} \approx A \mathrel{\vert\!\sim} B$. □

Proposition B.8. *If $\mathcal{D}$ is a deciding knowledge base for an entailment $\mathcal{K} \approx A \mathrel{\vert\!\sim} B$ and we have $\mathcal{D} \approx A \mathrel{\vert\!\sim} B$, then*

$$\mathcal{J}_W(\mathcal{D}, A \mathrel{\vert\!\sim} B) \subseteq \mathcal{J}_W(\mathcal{K}, A \mathrel{\vert\!\sim} B).$$

Proof. Since $\mathcal{D}$ is deciding, we have $\mathcal{D} \subseteq \mathcal{E}_A^{\mathcal{K}} \cup R^{-}(\mathcal{K}, A)$ and $\overline{\mathcal{D}} \models A \rightarrow B$. Every $\mathcal{J} \in \mathcal{J}_W(\mathcal{D}, A \mathrel{\vert\!\sim} B)$ has $\mathcal{J} \subseteq \mathcal{D}$ and $\overline{\mathcal{J}} \models A \rightarrow B$. Clearly every $\mathcal{J}$ is also minimal. Therefore $\mathcal{J} \in \mathcal{J}_W(\mathcal{K}, A \mathrel{\vert\!\sim} B)$. □

Our results suggest that these theorems hold for weak justification of Relevant Closure insofar as the properties for rationality apply to Relevant Closure itself. This provides support not just for our adaptation of weak justification to Relevant Closure but also for the properties themselves as the fact that they can be used even for the intuitive analysis of entailment formalisms that are not rational suggests that they express abstract and generalisable qualities of weak justification.

## B.2 Weak Justification of $A \mathrel{\vert\!\sim} B$ for $\mathrm{br}^{\mathcal{K}}(A) = \infty$

The case of $\mathrm{br}^{\mathcal{K}}(A) = \infty$ for an entailment $\mathcal{K} \approx A \mathrel{\vert\!\sim} B$ serves as an interesting boundary case for defeasible justification (and for that matter defeasible entailment). This occurs where the antecedent is exceptional even for the final element $\mathcal{E}_\infty^{\mathcal{K}}$ of the exceptionality sequence. In other words, the antecedent of the query contradicts classical information in the knowledge base. The following example illustrates this idea (note that we return to using $\approx$ to represent Rational Closure):

*Example B.9.* Consider the antecedent $p \wedge \neg b$ for $\mathcal{K} = \{p \rightarrow b\}$. We have such results as $\mathcal{K} \approx p \wedge \neg b \mathrel{\vert\!\sim} \bot$: since $p \wedge \neg b$ is a classical contradiction with respect to $\mathcal{K}$, it defeasibly implies every formula in $\mathcal{L}$ including contradictions (i.e. $\bot$).

This result above is not undesirable since this corresponds exactly to the behaviour of contradictions in the classical case and information in $\mathcal{E}_\infty^{\mathcal{K}}$ encodes classical assertions. However, there is some evidence that we ought to handle this as a special case for weak justification. Notice that each weak justification $\mathcal{J}$ of an entailment $\mathcal{K} \approx A \mathrel{\vert\!\sim} B$ with $\mathrm{br}^{\mathcal{K}}(A) = \infty$ has

$$\mathcal{J} \subseteq \mathcal{E}_\infty^{\mathcal{K}} \text{ and } \overline{\mathcal{J}} \models A \rightarrow B.$$

There are two possible cases here:

*Case* 1. $\overline{\mathcal{J}} \models A \rightarrow B$ but $\overline{\mathcal{J}} \not\models \neg A$.
*Case* 2. $\overline{\mathcal{J}} \models \neg A$ (for which necessarily $\overline{\mathcal{J}} \models A \rightarrow B$).

*Example B.10.* As in Example 4.9, consider $\mathcal{K} \approx p \mathrel{\vert\!\sim} x$ for

$$\mathcal{K} = \{p \mathrel{\vert\!\sim} b \wedge \neg f, p \mathrel{\vert\!\sim} x, p \mathrel{\vert\!\sim} \bot, b \mathrel{\vert\!\sim} f, b \mathrel{\vert\!\sim} \bot\}.$$

Every statement is in $\mathcal{E}_\infty^{\mathcal{K}}$, i.e. $\mathcal{K} = \mathcal{E}_\infty^{\mathcal{K}}$. Case 1 is for example represented by the weak justification $\{p \mathrel{\vert\!\sim} x\}$ while Case 2 is represented by weak justifications such as $\{p \mathrel{\vert\!\sim} \bot\}$ or $\{p \mathrel{\vert\!\sim} b \wedge \neg f, b \mathrel{\vert\!\sim} f\}$.

Of these two cases, it seems that Case 2 may necessitate special handling specifically due to justifications such as

$$\mathcal{J} = \{p \mathrel{\vert\!\sim} b \wedge \neg f, b \mathrel{\vert\!\sim} f\}.$$

In a defeasible context, $\mathcal{J}$ does not amount to sufficient reason to conclude $\neg p$ given the antecedent $p$ and the statements in $\mathcal{J}$ appear in $\mathcal{E}_\infty^{\mathcal{K}}$ because of other statements, namely $p \mathrel{\vert\!\sim} \bot$ and $b \mathrel{\vert\!\sim} \bot$. As we noted in Example 4.9, a consequence of this is that the propositions in Section 4 do not hold when the antecedent has base rank $\infty$.

A solution that seems logical is to require that all weak justifications of the Case 2 kind are such that every statement within the justification has an antecedent that is exceptional with respect to the justification:

$$\overline{\mathcal{J}} \models \neg C \text{ for all } C \mathrel{\vert\!\sim} D \in \mathcal{J}$$

where $\mathcal{J}$ is a candidate Case 2 weak justification. In principle, this should ensure that weak justifications of this kind do not contain defeasible statements. However, such weak justifications may not even exist due to the minimality condition:

*Example B.11.* Consider $\mathcal{K} \approx a \mid\sim \bot$ for

$$\mathcal{K} = \{a \mid\sim b \wedge c, b \mid\sim \neg c \wedge d, d \mid\sim \bot\} .$$

The sole (Case 2) weak justification is $\mathcal{J} = \{a \mid\sim b \wedge c, b \mid\sim \neg c \wedge d\}$ for which $\overline{\mathcal{J}} \models \neg b$.

Perhaps the existing requirement of minimality does not make sense in this specific circumstance, and perhaps we should be imposing minimality on the sets *after* requiring that $\overline{\mathcal{J}} \models \neg C$ for all $C \mid\sim D \in \mathcal{J}$. Otherwise, it may make sense—especially given the presumable difficulty of computing such knowledge bases—to simply accept the boundary case as it is and that results such as Propositions 4.4 and 4.5 may have special cases for $\text{br}^{\mathcal{K}}(A) = \infty$ of a query $A \mid\sim B$. We have not explored the matter any further, although we do note that Proposition 4.4 (and consequently Proposition 4.5) has a proof for $\text{br}^{\mathcal{K}}(A) = \infty$ after making the adjustment to weak justification described here.